Applying Web Mining and Sentiment Analysis to Assess Tourists Review on Batu City Tourist Destination

1st Aisyah Larasati
Department of Industrial Engineering
Universitas Negeri Malang
Malang, Indonesia
aisyah.larasati.ft@um.ac.id

2nd Joko Sayono
Department of History
Universitas Negeri Malang
Malang, Indonesia
joko.sayono.fis@um.ac.id

3rd Agus Purnomo
Department of Social Studies
Universitas Negeri Malang
Malang, Indonesia
agus.purnomo.fis@um.ac.id

4th Effendi Mohamad
Faculty of Manufacturing Engineering
University Teknikal Malaysia Melaka
Malaysia
effendi@utem.edu.my

5th Muhammad Farhan
Department of Industrial Engineering
Universitas Negeri Malang
Malang, Indonesia
muhfarhanaan11@gmail.com

6th Puji Rahmawati
Department of Industrial Engineering
Universitas Negeri Malang
Malang, Indonesia
pujirahma147@gmail.com

***Abstract*— Tourist attractions are one of the destinations for the community to eliminate fatigue and to function as a self-entertainment. Lots of favorite cities chosen by the community as a tourist destination, one of which is Batu City, East Java. This city has a variety of tourist sites, which can be categorized into two types, namely natural and artificial. Thousands of people come every year to spend their holidays in the city of Batu. For this reason, this study aims to apply web mining and sentiment analysis to determine people's sentiment or perspective on tourist attractions in Batu through the reviews found on the website. The web mining method is used to retrieve information from the travel website. Then, the retrieve information is analyzed by using python programming language. There are eight tourist sites whose reviews were taken, with the total number of reviews are 4887 reviews. The results show that the tourists sites have a polarity value with a range between 0 to 1, which indicates that all the tourists' site gets a positive sentiment. In addition, the resulting subjectivity value is in the range of 0.4 to 0.6, which indicates that all reviews given by visitors are opinions or personal opinions. It implies that each tourist site has sufficiently met the needs of visitors, regarding the facilities and infrastructure that are provided.**

Keywords— Tourist destination, Web Mining, Sentiment Analysis, Python

I. Introduction

Changes in lifestyle and human behavior are now increasingly dense and saturating. This condition affects on the choice of tourism destinations that are people visited as the place to eliminate their boredom from work and rountine daily life. Visiting tourist destinations is the activity that aims to refresh physically and psychologically to create a more positive mind for the body. Many cities have provided an interesting place as their tourist destination, one of which is Batu City. This is a small city next to Malang City and located in East Java Province, Indonesia. The city is famous for its nickname "tourist city" because it has a lot of potential natural beauty visited by domestic and foreign tourists. The object in Batu City is also diverse, from natural objects to artificial objects. The development of tourism in Batu City grows rapidly and are well responded by the surrounding community. Based on data in 2018, the regional income reached around 100 billion rupiahs and regional economic growth reached 8.3% [1].

Batu city has a big opportunity for tourism businesses to maintain their existence. The development carried out includes improving the quality of services, adding facilities and infrastructure that support the tourist object. On the other hand, many parties take the initiative to open new tourist objects by offering more attractive facilities and activities. This condition may increase the number of tourist objects in Batu City. A large number of objects that competes between the object to maintain the tourist market in order to obtain high income from tourist visits. Thus, evaluating and comparing the service quality of each tourism object becomes important [2].

Information about tourist preferences or opinions is now very easily accessible to the public and widely available in various website reviews. A large number of data reviews from tourists expresses the difficulty of visitors or the excitement of visitors during their visit are available online. To get useful information based on tourist responses, this study conducting a sentiment analysis, web mining and python programming. The advantages of web mining are to extract information on the website quickly and automatically [3]. Web mining also successful in the amount of voluminous data, so the new knowledge can be discovered to understand the patterns of customer behavior to support decision making [4].

The data processing method uses sentiment analysis. The benefit of sentiment analysis is to determine the preferences and emotions of the reviewer. Moreover, this method is reliable for classified words and ranks the popularity of tourist destinations [5]. The goal of this research is to assess tourist reviews in Batu City, both natural and artificial tourism objects using web mining and sentiment analysis methods. This method is more efficient to be applied to a large amount of data compared to conventional survey methods that use statistical techniques as data processing. The output of the analysis is considered more accurate since this method based on machine learning techniques, so it can determine visitor preferences based on wordlist frequency

978-1-7281-8131-8/20/$31.00 ©2020 IEEE

results, polarity values, and subjectivity values. The sentiment analysis is carried out with the *Python* programming language and *Anaconda* software. It is expected that this research can help tourism site managers to to make improvements to the facilities and infrastructure at their tourist sites more effectively.

II. RESEARCH METHODOLOGY

The process flow of this study is shown in Fig. 1. The data used in this study is collected from the traces of visitor reviews written on the tourist website of eight famous tourist destinations in Batu City. Overall, the total reviews taken were 4.887 reviews. The percentage number of reviews collected from each tourist site is shown in Fig. 2.

Web scrapping is used to perfom data collection. Web scraping is a web mining tool that focuses on retrieving data and information from the web automatically [6]. The process of web scraping initiated with determining the URL of web pages. The web pages used in this case are the web pages that contain reviews of tourist attraction in Batu City. Afterwards, the process of collecting data contained in URL of web page is carried out through an application such as web scraper by parsing the HTML elements of that web pages to extract the reviews automatically. Then, the data obtained through scraping process is used as input data to implement the analysis process such as data mining techniques, including sentiment analysis. This analysis aims to gain new insights or knowledge based on the data.

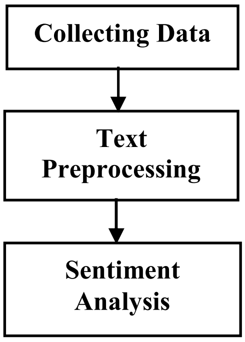

Fig. 1. Flowchart Project

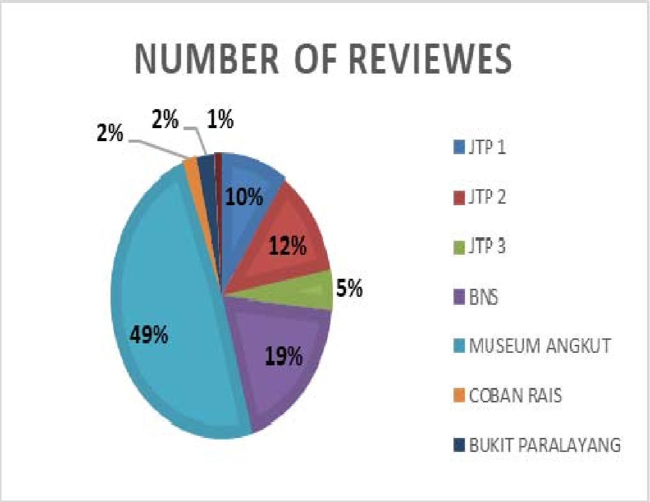

Fig. 2. Percentage Number of Tourist Attraction Reviews

The attraction reviews used is a type of natural and artificial tourism. The list of tourist sites used in the study can be seen in Table I. After visitor reviews are taken using a web scraper, then the Text Preprocessing is performed.

Text preprocessing is a stage that aims to prepare the review data that has been obtained, so that it can be processed in sentiment analysis. The steps taken in text preprocessing can be explain as follows.

- Tokenize, aims to break up sentences into separate words. For example, “This tourist site is very interesting and have a beautiful scenery”, this sentence breaks up to 10 words.
- Stopwords filter, aims to delete words that have no meaning (this, is, a, the, in, etc.). For example, the sentence in the previous section has 10 words, after doing the stopwords filter, the sentence change into “tourist site very interesting have beautiful scenery” and now only have 7 words.
- Stemming, aims to change words into basic forms of words. For example, the sentence in the previous section doing a stemming process, then the sentence change into “tourist site very interest have beautiful scenery”. The words “interesting” is a change to the basic form “interest”.

The words in the text preprocessing is used as an attribute for conducting sentiment analysis, so that it can classify customer reviews into positive, negative, or neutral sentiments. Sentiment analysis is performed using the help of python programming. The tools used to assist in sentiment analysis are Jupyter Notebook, which can be obtained at Anaconda.com. After opening a new sheet on Jupyter Notebook, the initial step is to install TextBlob and Python Sastrawi. TextBlob is one of the python libraries used to process textual data and can be used to help Natural Language Processing (NLP). Meanwhile, Python Sastrawi is also one of the python libraries that can be used to stem the Indonesian language text. The TextBlob library can be installed by pip installing textblob, and for literary python, it can be done by pip installing Sastrawi. Next, import the required modules and create variables from the literary python module for the Indonesian stemming process. Then, to read the review data that has been obtained previously, this study uses the Pandas. Pandas are used to help cleaning up raw data into a better form for analysis. After the data has been successfully read, the next step is to delete the unneeded portion of the column. Next, do the stemming process based on the previously created variable. Then, Indonesian-language textual data is translated into English, and the final step is to do sentiment analysis.

TABLE I. LIST OF TOURIST SITE

No.	Tourist Site
1.	Jatim Park 1 (JTP 1)
2.	Jatim Park 2 (JTP 2)
3.	Jatim Park 3 (JTP 3)
4.	Museum Angkut
5.	Batu Night Spectaculer (BNS)
6.	Bukit Paralayang
7.	Coban Rais
8.	Rabbit Park

The sentiment analysis results can be categorized into two forms, namely polarity, and subjectivity. In polarity, the value between the range 0-1 is a positive sentiment, if the generated value is 0, then it implies neutral sentiment. If the value is between -1 and 0 range, then the data is negative sentiment. On the other hand, subjectivity is used to see whether the reviews given are fact or opinion. If it's a fact, then the resulting value is 0. If it is an opinion, then the value produced is between 0 and 1. So, from the result of the sentiment analysis, it informs whether the visitor's sentiments towards attractions in Batu City are positive, neutral, or negative.

III. RESULT AND DISCUSSION

Python programming is used to define and analyze sentiment analysis in the tourist destinations in Batu, Malang. Eight attractions were used to do the analysis. From the data processing that is performed, the polarity and subjectivity results obtained for each tourist attraction can be seen in the Table II, with P for Polarity, and S for Subjectivity.

As the results shown in Table II, it can be seen that all tourist objects have a polarity value between 0-1 which means the reviews for all objects have a positive sentiment. As for subjectivity, all of the attractions studied have a range of values between 0 to 1, which means the review given by visitors is an opinion or personal opinion. Also, from the analysis, wordlist are obtained to show the frequency of eachword.

A. Jatim Park 1

The results of the frequency of words in a review of Jatim Park 1 is shown in Table III. Table III shows the top 15 words that have the most frequencies with the total reviews from the website is 412.

TABLE II. SENTIMENT ANALYSIS RESULT: POLARITY & SUBJECTIVITY

No.	Artificial Tourist Attraction	P	S
1	Jatim Park 1	0.3324	0.6546
2	Jatim Park 2	0.3324	0.6546
3	Jatim Park 3	0.1583	0.4416
4	Museum Angkut	0.6210	0.5418
5	Batu Night Spectacular (BNS)	0.2441	0.5188
7	Bukit Parayalang	0.3433	0.6188
7	Coban Rais	0.2200	0.6655
8	Rabbit Park	0.2891	0.6335

TABLE III. WORDLIST OCCURRENCE OF JATIM PARK 1

No	Word	Total Occurrence	Document Occurrence
1	good	72	62
2	interesting	48	44
3	complete	42	38
4	exciting	40	30
5	fun	38	34
6	large	30	30
7	educative	27	23
8	cheap	27	25
9	not bad	25	23
10	cool	21	21
11	clean	19	17
12	tour_educational	16	16
13	adrenalin_pumped	15	15
14	crowded	15	14
15	children_suitable	14	14

The results show that most of the responses from visitors imply positive responses. This can be seen from the word used in the reviews, such as “good”, “interesting”, “complete”, “exciting”, “fun”, as well as "suitable with children". These word express that tourists feel happy during their visit in Jatim Park 1. For the place, visitors perceive that Jatim Park 1 has a large area, cool, clean, and crowded place. Based on these results, this study concludes that visitors feel quite satisfied with what has been provided by Jatim Park 1 attraction as well as they have a nice experience during their visit in Jatim Park 1.

B. Jatim Park 2

The results of the frequency of words in a review of Jatim Park 2 is shown in the Table IV.

Table IV shows the top 15 words that have the most frequencies with the total reviews from website is 530. This results show that most of vistors deliver a positive responses. Tourist perceive that Jatim Park 2 has a good place with the large area, clean, complete, and comfortable place. Also, it shows from the word occurrence that Jatim Park 2 has a cheap ticket price, however some other tourists also say its expensive. But in general, tourist gives a positive response because this tourist attraction is clean, interesting, fun, manicured, and comfortable. Jatim Park 3

The results of the frequency of words in a review of Jatim Park 3, shown in the Table V.

TABLE IV. WORDLIST OCCURRENCE OF JATIM PARK 2

No	Word	Total Occurrence	Document Occurrence
1	good	130	108
2	complete	79	73
3	Large	75	72
4	clean	74	72
5	interesting	54	50
6	fun	53	52
7	impressive	26	22
8	Like it	26	24
9	manicured	26	23
10	best	25	25
11	cool	23	22
12	comfortable	21	19
13	cheaps	20	19
14	exciting	20	18
15	expensive	17	17

TABLE V. WORDLIST OCCURRENCE OF JATIM PARK 3

No	Word	Total Occurrence	Document Occurrence
1	legend	53	35
2	large	43	37
3	great	43	30
4	interesting	34	27
5	good	29	25
6	like it	19	15
7	fun	18	14
8	exciting	17	9
9	complete	16	16
10	impressive	16	11
11	recommend	14	11
12	not bad	9	9
13	like_children	9	8
14	clean	9	8
15	expensive	9	8

Table V shows the top 15 words that have the most frequencies with the total reviews from website is 195. Just like Jatim Park 1 and Jatim Park 2, Jatim Park 3 also receives a positive response from the reviews given by tourists on the website. Tourists imply that JTP 3 ia a legend place, has large and clean area. In addition, based on the wordlist occurrence, it show that Jatim Park 3 is an interesting place, fun, complete, and children like it. Thus, it conclude that the tourists are quite satisfied with what is provided by the tourist attractions.

C. Museum Angkut

The results of the frequency of words in a review of Museum Angkut is shown in Table VI. The total reviews of Museum Angkut obtained from website is 2084. Table VI shows the top 15 word with the highest occurrence. This results imply that Museum angkut also receives a fairly positive comments from the tourists. The visitors say that Museum Angkut has a great and huge place, unique, and clean. Moreover, some tourists also say museum angkut is one of the tourist attractions that worth to visit, recommend, awesome, and perfect place to visit. Overall, Museum Angkut provides services that are satisfied the tourists.

D. Batu Night Spectacular (BNS)

The results of the frequency of words in a review of Batu Night Specatcular is shown in Table VII.

Table VII shows that in general Batu Night Spectacular (BNS) receives a positive review from the reviews in the website. This table shows 15 top word with highest occurrence. There are 817 reviews obtained from the website. Tourists perceive that Batu Night Spectacular is a nice place, interesting place, beautiful place, and large place to spend their time. However, some tourists say that ticket price is quite expensive, but others also say the ticket price is quite cheap. Thus, from the reviews collected from website, it describes that BNS is a recommended place to spend vacation time because it provides a good place, interesting game rides and complete infrastructures.

E. Bukit Paralayang

The results of the frequency of words in a review of Bukit Paralayang is shown in Table VIII. The total number of reviews collected from the website is 106. The visitors' reviews show that most of visitors deliver a positive response to Bukit Paralayang. The visitors consider this place as an attractive tourist destination that has a beautiful-night-view with a mountain scenery, where the view presented makes visitors enjoy the trip.

F. Coban Rais

The results of the frequency of words in a review of Coban Rais is shown Table IX. The total number of reviews is 90.

TABLE VI. WORDLIST OCCURRENCE OF MUSEUM ANGKUT

No	Word	Total Occurrence	Document Occurrence
1	good	692	542
2	interest	555	426
3	fun	495	422
4	great_place	303	182
5	awesome	301	233
6	attract	246	207
7	unique	232	201
8	perfect	188	171
9	worth_visit	174	156
10	huge	147	131
11	crowd	98	85
12	recommend	89	80
13	clean	78	76
14	cool	54	43
15	impress	40	38

TABLE VII. TABLE 7. WORDLIST OCCURRENCE OF BATU NIGHT SPECTACULAR

No	Word	Total Occurrence	Document Occurrence
1	nice	211	178
2	interesting	133	112
3	enjoy	92	84
4	exciting	79	73
5	cheap	76	74
6	cool	75	72
7	beautiful	71	67
8	fun	69	65
9	not bad	68	64
10	crowded	60	57
11	like it	46	42
12	expensive	44	42
13	complete	41	40
14	large	29	27
15	interesting_game rides	25	25

TABLE VIII. WORDLIST OCCURRENCE OF BUKIT PARALAYANG

No	Word	Total Occurrence	Document Occurrence
1	paralayang	75	60
2	batu	58	48
3	city	47	40
4	view	35	34
5	night	34	27
6	malang	29	21
7	batu_city	28	27
8	mountain	25	18
9	tourism	23	17
10	location	22	16
11	beautiful	21	19
12	road	19	16
13	good	17	15
14	flying	17	12
15	enjoy	16	16

TABLE IX. WORDLIST OCCURRENCE OF COBAN RAIS

No	Word	Total Occurrence	Document Occurrence
1	Entrance	34	18
2	child	33	7
3	taxibike	26	17
4	tourism	24	15
5	photo spot	23	18
6	beautiful	22	15
7	walk	20	12
8	rp	20	5
9	nature	18	13
10	visitors	17	9
11	trip	16	10
12	feet	15	14
13	people	15	8
14	ticket	15	10
15	children	14	3

TABLE X. WORDLIST OCCURRENCE OF RABBIT PARK

No	Word	Total Occurrence	Document Occurrence
1	rabbit	64	30
2	child	32	18
3	park	26	18
4	rabbit_park	14	12
5	children	12	11
6	tourism	12	6
7	photo	10	8
8	those rabbits	10	9
9	entrance	10	9
10	nice	9	7
11	rabbit	9	4
12	ticket	9	8
13	area	8	5
14	batu	8	8
15	location	8	6

Table IX shows the frequency of words Occurrence based on visitor reviews in Coban Rais. Similar to popular emotions in Coban Rondo, visitors were fascinated to the beautiful scenery of Coban Rais. Most of them visited Coban Rais to experience the water fall and take some pictures before ending the trip. Since the natural beauty spoils the visitor, it makes Coban Rais become the recommendation destinations for spending time with family and that's why Coban Rais is full of visitors at all times if holiday comes.

G. Rabbit Park

The results of the frequency of words in a review of Rabbit Park, shown in Table X, with the total number of reviews conducted from website is 45. Table X shows the frequency of words that often appears in the results of tourism reviews of Rabbit Park. The word frequency list above shows that most reviewers responded positively to Rabbit Park. Visitor preferences on the Rabbit Park tourism object are recommended as a suitable location for children's because there are many rabbit animals are provided that can be used as children's education.

Rabbit Park also has many great photo spots that can enhance the aesthetic value of this object. Thus, the majority of visitor reviews give a good rating on the Rabbit Park tourism object.

From all the reviews of tourist that shows in the table for each tourists' attraction, it can be conclude that 10 tourist attraction in Batu get a positive response from the tourists. It is also shown by the value of polarity which shows that all attractions are in the range of values from 0-1 which means the object has a positive sentiment. For subjectivity value it describes that all the reviews for tourist attraction are a personal opinion.

There some research related to this paper, such as [7] doing research that use sentiment analysis to identify their review quality, review characteristics, and review sentiment from online hotel reviews using classification techniques such as decision tree, logistic regression, random forest, and support vector machine. And also, research [8] that applied a sentiment analysis method to analyze traveler's online reviews to know about how the response from travelers about tourism destination, and the result show services that have the lowest quality value are transportation. In the field of airport services [9], their research try to assessing the level of services perceived by airport customer from the data that was collected from the blog by using text extracting software. Also, some of paper that used sentiment analysis in their research are [10] used naïve bayes and support vector machine as the algorithm, [11] used text mining to get information from people feedback in twitter, [12] used integrated framework such as data crawler and kernel classification to gain insight about hotel reviews and rating, [13] used sentiment analysis to identify and analyze emotion and opinions from the residents' or tourists' review, and [14] used twitter to collect the reviews to get the insight about tourist sentiment for the local city.

In this paper we used web mining method and python programming to get the insight about the sentiment of tourists about tourists' attraction in Batu, Malang. Most of the results shows each tourist attraction have a positive sentiment. It is quite difficult to analyze a negative sentence from the reviews, because sometimes negative sentiments are contained implisitly in a sentence. So, the sentence may show bias sentiment that it can also affect the results obtained from the analysis. This is also supported by the paper [14] that doing a research on assessing and reviewinge different sentiment analysis approaches. Their results show that most of sentiment analysis used a machine learning approach to process their data doesn't work so well when classifying negative or positive sentences. However, the machine learning works better on positive sentences than on a negative one. Furthermore, the most important problem when doing a sentiment analysis is to discover implicit aspects from the sentences. So, there are quite a lot of possibilities if the analysis carried out does not produce a quite accurate results, because many factors may affect the process. Thus, to overcome this problem, it is better if the reviews on each tourist attraction are made in a form of dictionary word when they are processed at the stopwords stage. In addition, each review should be adjusted accordingly to obtain more accurate sentiment results since each object has different characteristics. Whereas in the web mining process, it is better to use software such as R Studio, because it can extract reviews automatically, so that the process to extract information from the website is going to be easier and require less time.

IV. CONCLUSION

The easier access to the internet makes it easier for people to give their views specifically on online media such as websites. There are many websites devoted to various fields, for example for business, politics, technology, and one of them is a website for tourism. To find out the response given by visitors on eight tourist destinations, this study extract the visitors review the travel website. The results show all tourists attractions get a positive sentiment, which can be seen from the value of polarity and the wordlist occurrences. For subjectivity, it shows that all of the tourists' attraction reviews are personal opinion. The result implies that eight tourist attractions: Jatim Park 1, Jatim Park 2, Jatim Park 3, Museum Angkut, Batu Night Spectaculer, Rabbit Park, Bukit Paralayang, and Coban Rais are received good enough comments from tourists. These results indicate that the tourist destinations have provided sufficient facilities and other needs of visitors. However, the manager of the tourist destinations must improve the quality of services continuously since many tourists perceived that the ticket prices is quite expensive.

ACKNOWLEDGMENT

The authors would like to acknowledge funding support from Universitas Negeri Malang (UM), which has funded this study through PNBP – UM Research Grant 2020 under contract No.4.3.675/UN32.14.1/LT/2020.

REFERENCES

[1] K. Abdul, "Tourism And Development: Land Acquisition, Achievement Of Investment And Cultural Change (Case Study Tourism Industry Development In Batu City, Indonesia)," *GTG*, vol. 21, no. 1, p. 253, 2018, doi: 10.30892/gtg.21120-285.

[2] A. Tjahyanto and B. Sisephaputra, "The Utilization of Filter on Object-based Opinion Mining in Tourism Product Reviews," *Procedia Computer Science*, vol. 124, pp. 38–45, 2017, doi: 10.1016/j.procs.2017.12.127.

[3] S. Mowla, I. Bedi, and N. P. Shetty, "A Study on Web Mining Tools and Techniques," *Journal of Engineering and Applied Sciences*, p. 10, 2018.

[4] G. T. Wei, S. Kho, W. Husain, and Z. Zainol, "A Study of Customer Behaviour Through Web Mining," *J. Inf. Sci. Comput. Technol.*, vol. 2, no. 1, pp. 103–107, 2015.

[5] A. Rajan, "Sentiment Analysis on Customer Reviews in Tourism - A Text Mining Approach," 2015.

[6] C. Slamet, R. Andrian, D. S. Maylawati, Suhendar, W. Darmalaksana, and M. A. Ramdhani, "Web Scraping and Naïve Bayes Classification for Job Search Engine," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 288, no. 1, 2018, doi: 10.1088/1757-899X/288/1/012038

[7] P. J. Lee, Y. H. Hu, and K. T. Lu, "Assessing the helpfulness of online hotel reviews: A classification-based approach," *Telemat. Informatics*, vol. 35, no. 2, pp. 436–445, 2018.

[8] K. Kim, O. joung Park, S. Yun, and H. Yun, "What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management," *Technol. Forecast. Soc. Change*, vol. 123, pp. 362–369, 2017

[9] S. Gitto and P. Mancuso, "Improving airport services using sentiment analysis of the websites," *Tour. Manag. Perspect.*, vol. 22, pp. 132–136, 2017

[10] J. C. L. Menchavez and K. J. P. Espinosa, "Fun in The Philippines : Automatic Identification and Sentiment Analysis of Tourism-related Tweets," pp. 660–667, 2015.

[11] V. Ramanathan, "Twitter Text Mining for Sentiment Analysis on People ' s Feedback about Oman Tourism," *2019 4th MEC Int. Conf. Big Data Smart City*, pp. 1–5, 2019.

[12] Y. Chang, C. Ku, and C. Chen, "International Journal of Information Management Social media analytics : Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor," *Int. J. Inf. Manage.*, no. April, pp. 0–1, 2017.

[13] S. Gao, J. Hao, and Y. Fu, "The application and comparison of web services for sentiment analysis in tourism," *2015 12th Int. Conf. Serv. Syst. Serv. Manag. ICSSSM 2015*, 2015.

[14] A. R. Alaei, S. Becken, and B. Stantic, "Sentiment Analysis in Tourism: Capitalizing on Big Data," *J. Travel Res.*, vol. 58, no. 2, pp. 175–191, 2019