Jurnal E-KOMTEK (Elektro-Komputer-Teknik)
Vol. 5, No. 2 (2021) pp. 284-297
https://jurnal.politeknik-kebumen.ac.id/index.php/E-KOMTEK
p-ISSN : 2580-3719 e-ISSN : 2622-3066

Big Data Analytics to Analyze Sentiment, Emotions, and Perceptions of Travelers (Case Study: Tourism Destination in Purwokerto Indonesia)

Siti Khomsah[1], **Rima Dias Ramadhani**[2], **Sena Wijayanto**[3]
[1,2]Department of Data Science, Institut Teknologi Telkom Purwokerto, Indonesia, 53147
[3]Department of Information System, Institut Teknologi Telkom Purwokerto, Indonesia, 53147

siti@ittelkom-pwt.ac.id

https://doi.org/10.37339/e-komtek.v5i2.791

Published by Politeknik Dharma Patria Kebumen

Artikel Info
Submitted:
10-12-2021
Revised:
24-12-2021
Accepted:
24-12-2021
Online first :
30-12-2021

Abstract

Big data analytics can extract travelers' sentiment, emotions, and experiences from their internet opinions. This study analyzes sentiment, emotion, and traveler experiences at eight tourism destinations in Purwokerto Central Java, Indonesia. The methods are lexicon using NCR vocabulary(EmoLex) and word cloud analysis. The results show visitors generally have a positive sentiment. The five destinations with high positive sentiment are the Village (91%), Lokawisata Baturaden(81%), Baturaden Forest (79%), Limpa Kuwus (78%), and Taman Andang(.77%). In comparison, other destinations achieve positive sentiment under 70%. Only a few visitors give negative sentiment to all tourism destinations. The emotion of visitors stands out in Joy and Trust. NRC revealed sadness dan anger emotion but only about 20%. Cloud analysis does not reveal a distinguish keyword because the word feature still contained noise such as conjunction, adverb, and the name of the sites. Further research must consider other text preprocessing to handle noises.

Keywords: Big data analytics, Tourism, Sentiment analysis, Emotion analysis, Word cloud

Abstrak

Big data anaytcs dapat mengekstrak sentimen, emosi, dan pengalaman wisatawan dari opini mereka di internet. Penelitian ini bertujuan untuk menganalisis sentimen, emosi, dan pengalaman wisatawan di delapan destinasi wisata di Purwokerto Jawa Tengah Indonesia. Metode yang digunakan adalah leksikon menggunakan kamus NCR (EmoLex) dan word cloud. Hasil penelitian menunjukkan pengunjung umumnya memiliki sentimen positif. Lima destinasi dengan sentimen positif tinggi adalah The Village (91%), Lokawisata Baturaden (81%), Baturaden Forest(79%), Limpa Kuwus (78%), dan Taman Andang (0,77%). Sedangkan destinasi lainnya mencapai sentimen positif di bawah 70%. Hanya sedikit pengunjung yang memberikan sentimen negatif ke semua destinasi wisata. Emosi pengunjung menonjol dalam Joy (senang) and Trust(percaya). NRC mengungkapkan emosi sedih dan marah tetapi hanya sekitar 20%. Analisis cloud tidak mengungkapkan kata kunci yang membedakan karena fitur kata masih mengandung noise seperti kata sambung, kata keterangan, dan nama tempat. Penelitian selanjutnya harus mempertimbangkan preprocessing teks yang baik untuk menangani noise.

Kata-kata kunci*: Analisis data besar, Pariwisata, Analisis sentimen, analisis emosi, Awan kata*

1. Introduction

Big data analytics is a technique for analyzing and extracting large and complex information. Big data have characteristics of 3V, volume, velocity, and variety [1]. Management and analysis of large amounts of data cannot use the traditional's approach [1].Big data analytics includes capturing, storing, accessing, and analyzing data to gain knowledge and use it for decision-making [1]. Big data is in various fields such as health, tourism, education, business, environment, government, industry, agriculture, and other domains. The utilization of big data analytics in various fields is an opportunity to gain knowledge based on data. Likewise, the application of big data analytics in tourism [2]. Tourism is related to tourism destinations, but it involves accommodation such as hotels, transportation, restaurants, handicrafts, and food industries [3]. In the digital era, almost all services related to tourism can be found on the internet and social media, such as Flickr, Google Maps, and Instagram. Travelers' opinions on the internet are found in large numbers in the text and images. We can easily find reviews of hotels, food, restaurants, or tourism destinations. The user's opinions can be lead to the choice of potential consumers. For example, users choose a hotel based on reviews from others' customer satisfaction and rating on Google Map Review [4][5]. Big data analytics can extract information from comments, images, or spatial data [6]. The gold information extracted from reviews is sentiments and emotions [7].

Several studies show that tourist experiences can be detected from reviews and pictures uploaded on social media or the internet. Most young and male tourists or travelers usually share their travel experiences on Instagram [4]. Travelers tell the story and atmosphere of the tourism site on Facebook to help other tourists to get the best travel trip [5]. Many methods are used to extract knowledge from the source text. Lexicon is used to reveal the sentiments and emotions of visitors to tourism sites. Sentiment analysis using the lexicon is useful for knowing the emotional condition of netizens [8]. Several studies generally use word clouds to find out the dominant words in a set of sentences with positive and negative sentiments [9], summarize conversations [10] and utilize word cloud analysis to get the most popular keyword on the dataset and to find the relationship between sentiment and reality [11]. A word cloud can also figure out important keywords trending among tourist reviews toward some destinations.

This study analyzes visitors' sentiments, emotions, and perceptions of tourism destinations in Purwokerto Central Java, Indonesia. We used a dataset from a Google Map review of some tourist sites in Purwokerto. We used word cloud to extract keywords that express the travelers' experiences toward tourist destinations.

2. Method

This section describes the methodology of this research strategy, begins with data crawling. The second step is data pre-processing. The third step is feature extraction, followed by applying sentiment and emotion analysis and word cloud analysis, shown in Figure 1. The detailed methods are also discussed in this section.

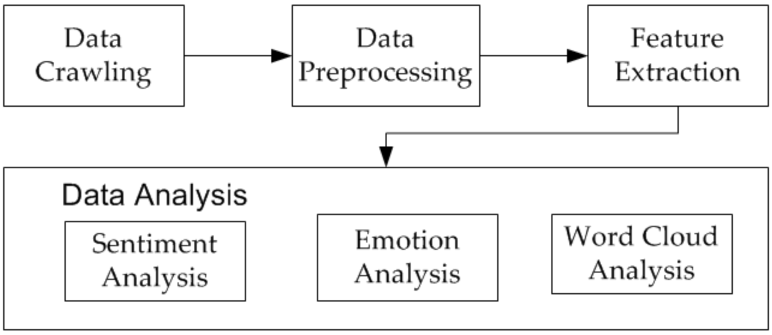

Figure 1. Steps of Research Method

a. Data Crawling

We harvest data from Google Map Review using the software webharvy (https://www.webharvy.com). Eight tourist destinations were chosen as the sample. The name of the destination and number of comments (reviews) from visitors are shown in Table 1.

Table 1. Dataset

Tourist Destination	Amount of Reviews
Baturaden Forest	291
Limpak Uwus	181
Lokawisata Baturaden	291
Museum BRI	271
Taman Andang	311
Curug Telu	311
Telaga Sunyi	371
The Village	181

Data are contained user names, ratings, and comments. In this research, we do not consider the ratings because, in those samples, we found inconsistency between ratings and comments. We only extract sentiment and emotion from data comments.

b. Data Preprocessing

Data are contained user names, ratings, and comments. In this research, we do not consider the ratings because, in those samples, we found inconsistency between ratings and comments. We only extract sentiment and emotion from data comments.

1) They are: deleting all characters other than the alphabet. All characters other than the alphabet, including numbers, are devoid of sentiment and emotion. Sentiments and emotions will appear only in words that have meaning. Numbers and symbols other than the alphabet are ignored.
2) Convert all capslock letters to lowercase. Standardization letters are important because the computer can not distinguish between uppercase and lowercase letters.
3) Stop Word Removing. If a word is found in the stop word list, it will be deleted. The stop word list uses the Sastrawi library.
4) Handling slang words. Slang word usually contains many repetitive characters. In an example, the word "siiiaaaappppp" (ready) will be transformed to "siap" (ready). "Waaaahhh"(very good) becomes "wah" (very good), and "maantuuulll" (very good) becomes "mantul" (very good). And also, a word with only one character will be deleted.
5) Then it should convert the slang words into KBBI (Indonesian vocabulary). For example, "guwee" (me) is converted to "aku"(me), "eloo"(you) is converted to "kamu"(you), "laen" is converted to "lain"(other), and so on. The researcher made a slang dictionary for this conversion process containing slang vocabulary.
6) Stemming is transforming the affixed words into the basic form of the word (stem). It is useful for uniforming words with the same meaning even though they have different suffixes. For example, the word "nikmat" (enjoyment)and the word "kenikmatan" (enjoyment) of them contain elements of positive sentiment. The two words will be converted into one basic word, namely "nikmat"(enjoyment), which also has a positive sentiment. Examples of other words are "kesulitan" (difficulty) and "sulit"(difficult) will be converted to "sulit" (difficult).

c. Feature Extraction and Data Analysis

The results of pre-processing are semi-structured. Patterns of sentiment and emotion are more apparent, but we need a method to extract sentiments and emotions from those semi-structured data. Sentiment extraction also labels sentiment polarity, namely positive and negative. While and labels for emotional polarity, namely anger (angry), disgust (disgust), sad (sadness), fear (fear), happy (joy), surprised (surprise), sure (trust) as shown in Table 2.

Table 2. Dataset

Polarity	Labels	Score
Positive	Positive Sentiment	1
Negative	Negative Sentiment	-1
Angry	Angry Emotion	0 - 1
Disgust	Disgust Emotion	0 – 1
Sadness	Sadness Emotion	0 – 1
Fear	Fear Emotion	0 – 1
Joy	Joy Emotion	0 – 1
Surprise	Surprise Emotion	0 – 1
Trust	Trust Emotion	0 - 1

All comments in the dataset are labeled automatically using the Lexicon method. The lexicon used in this study is the NRC EmoLex, the emotion lexicon by Saif Mohammad [13]. It lists the lexicon and its sentiments and emotions in 2 groups of sentiment types (positive and negative) and 8 groups, as shown in Table 2. According to our research needs, we did not include anticipation emotions from the NRC dictionary. Algorithmically, the NRC Dictionary is prepared in a table with a one-hot-encoded structure, as shown in Table 3 [13].

Table 3. Struktur NRC Emotion Leksikon

Keyword	Positive	Negative	Angry	Disgust	Sadness	Fear	Joy	Surprise	Trust
hancur(broken)	0	1	1	1	0	1	0	0	0
Keras(hard)	0	1	0	1	0	0	0	0	0
menangis(cry)	0	1	0	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…
Key-n	1	0	0	0	0	1	0	1	1

The code numbers 0 and 1 in Table 3 above show sentiment and emotion elements in keywords. The NRC dictionary in Table 3 is stored in memory space. It is well used to extract sentiments and emotions in the dataset. Extracting sentiments and emotions uses Lexicon, and the logical step is presented in the pseudocode below.

```
for y in dataset['Review']:
for teks in y.split():
if teks==positive
  f_sentimenpositif +=1
if teks==negative:
  f_sentimennegatif+=1
if teks==anger:
  f_anger+=1
```

```
if teks==fear:
  f_fear+=1
if teks==disgust:
  f_disgust+=1
if teks==sadness:
  f_sadness+=1
if teks==joy:
  f_joy+=1
if teks==surprise:
  f_surprise+=1
if teks==trust:
  f_trust+=1
totalemotion=max([f_anger,f_fear,f_disgust,f_sadness,f_surprise,f_joy,f_trust])
l_anger=f_anger/totalemotion
l_disgust=f_disgust/totalemotion
l_fear=f_fear/totalemotion
l_sadness=f_sadness/totalemotion
l_surprise=f_surprise/totalemotion
l_joy=f_joy/totalemotion
l_trust=f_trust/totalemotion
```

Word clouds analysis aims to understand the most frequent conversations indicated by the dominant words. Word cloud is a graphical representation of word frequency that gives greater importance to words that show more frequently in a source text. A Word cloud will portray words based on their frequency. The size of the word shows how important it is.

3. Result and Discussion

This section discusses and analyzes the output from extraction. We will discuss output, sentiment, emotion, and word clouds. Sentiment represents the polarity of opinion, both negative and positive. The sentiment is outlined into types of emotions. Word clouds represent the dominant keyword. The details will describe below.

a. Sentiment Analysis

The sentiments on the 8 tourism destinations are shown in Figure 2. Generally, commentators (travelers) show positive sentiments toward all tourism destinations. The percentage of positive sentiment is between 55% and 91%. The tourism site that received the highest positive sentiment is The Village with 91%. Lokawisata Baturaden, Baturaden Forest, Limpa Kuwus, and Taman Andang also get positive sentiment, but only around 77% to 81%. Meanwhile, others sites have positive sentiment under 70%, Telaga Sunyi, Museum BRI, and Curug Telu. Even though the destinations get a low positive sentiment, it does not mean that those sites have a high negative sentiment. As shown in Figure 2, negative sentiment for all tourism destinations is under 20%.

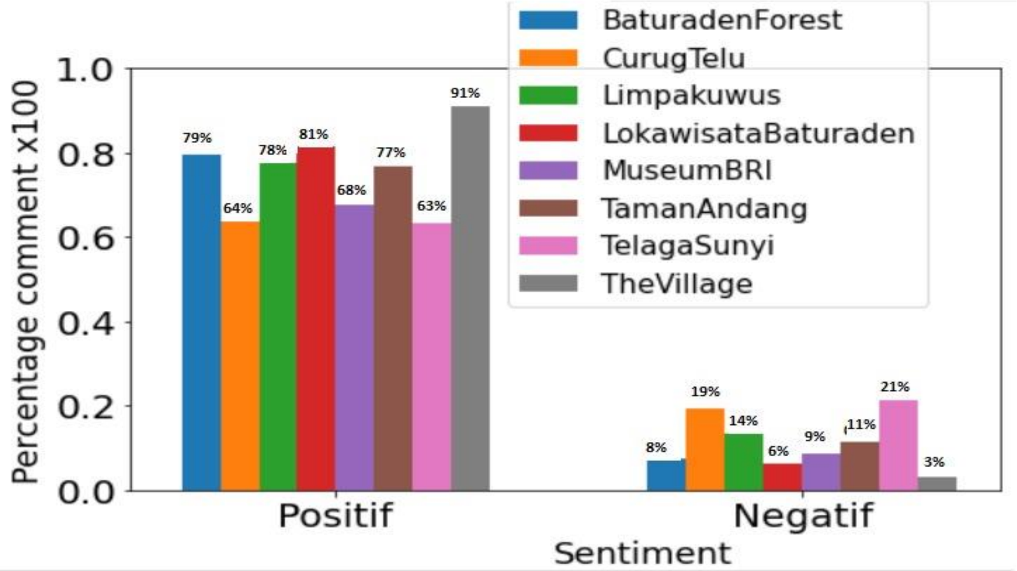

Figure 2. Sentiment Of Visitors Toward Tourism Sites

b. Emotion Analysis

There are two groups of emotions, positive emotion (joy, surprise, and trust) shown in Figure 3 and negative emotion (angry, sadness, fear, and disgust) shown in Figure 4. Mostly, eight tourism destinations make visitors joy and trust, in the range between 44% and 69%. The destination that presents the highest joy emotion is Taman Andang with 65%, whereas trust emotion is The Village with 69%.

Negative emotion toward eight destinations is quite low, below 30%. It describes the visitors got positive emotion while they have been spent time in each destination. Among the 8 destinations, Telaga Sunyi presented the most conspicuous negative emotions. Some visitors (28%) feel sad emotions while in Telaga Sunyi, and 22% of visitors experience anger.

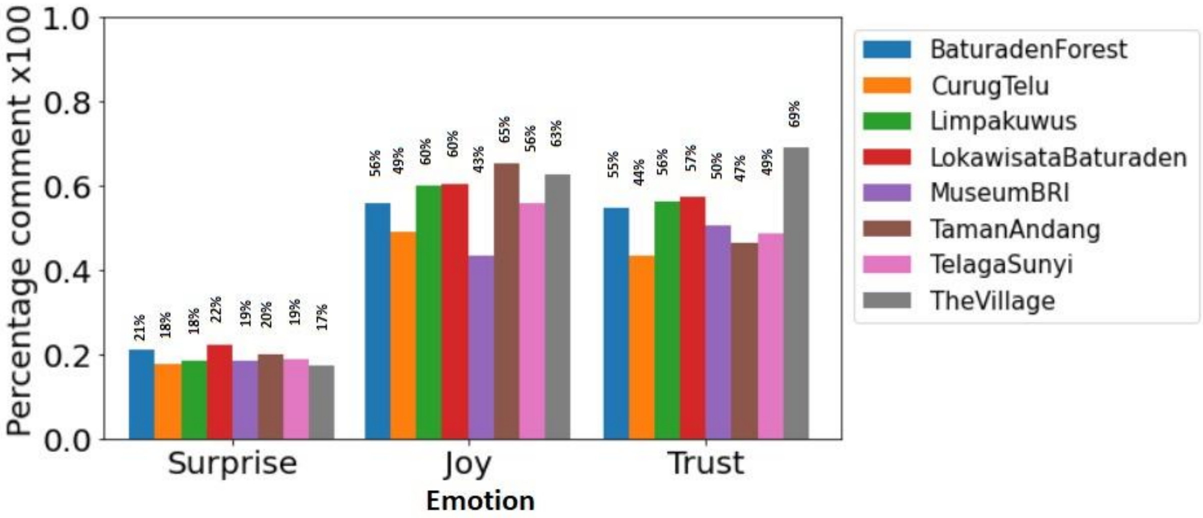

Figure 3. Negative Emotions Of Visitors Toward Tourism Sites

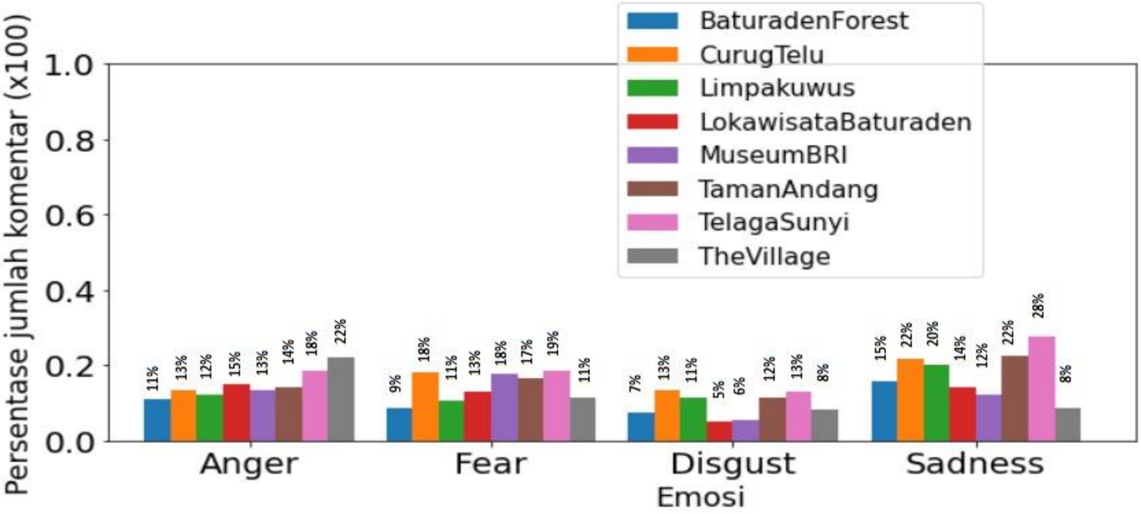

Figure 4. Negative Emotions Of Visitors Toward Tourism Sites

c. Word Cloud Analysis

Reviews of each of the 8 tourism sites will be presented in a word cloud. The word with the largest size shows visitors' perception of those places.

1) Word Cloud Baturaden Forest

Baturaden Forest is a natural forest at the foot of Slamet mountain. It is a favorite place for travelers. Many reviews about Baturaden Forest can be pictured as a word cloud, as shown in Figure 5. Those word cloud shows keywords such as “makan”(eat), “asli”(natural), “sejuk”(chill), “bagus”(mesmerizing), “pandang”(seeing), ”alam”(nature), “nyaman”(comfort). Noises words such as “sangat”(very)," "buat”(for)" are still quite a lot, if we ignore them, the cloud show that visitors consider Baturaden Forest as a place that chill, natural, nice place to eat, and sightseeing.

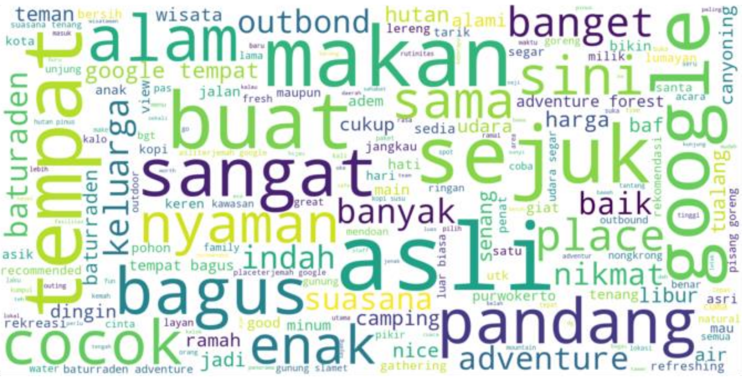

Figure 5. Word Cloud Baturaden Forest

2) Word Cloud Limpa Kuwus (The Limpa Kuwus)

Limpa Kuwus is a place like a natural forest that is managed well. Limpa Kuwus offers natural tourism. Word cloud in **Figure 6** shows the Limpakuwus revealed several unique keywords such as ”sejuk”(chill), “nyaman”(comfortable), “bagus”(beautiful), “jalan”(walk). This word cloud still contains many noise words (tempat/ place, buat/for, masuk/ entry, etc.). If we ignore noise’s word, the largest size is the word of chill (sejuk), visitors consider Limpa Kuwus as a cold place.

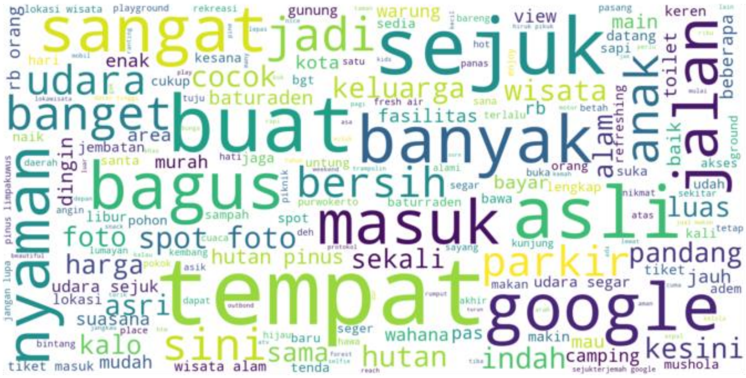

Figure 6. Word Cloud Limpa Kuwus

3) Word Cloud Lokawisata Baturaden (The Baturaden Tourist Site)

Lokawisata Baturaden is one of the large tourist destinations in Purwokerto. This place is a forested hill with various facilities. There are parks, restaurants, and walking tracks. Tourists can find many villas in Baturaden. Based on reviews of travelers, we can build a word cloud of Lokawisata Baturaden, as shown in **Figure 7.** It contains a lot of noise words such as “banyak”(many), “tempat”(place), “wisata(tour), where these words are not meaningful. If that noise is ignored, we will see the word with the largest size is nature(Alam). It shows that Baturaden Lokawisata is perceived as a nature tourism destination.

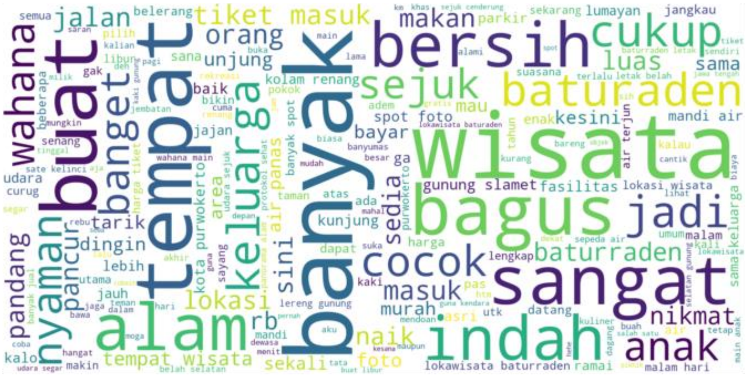

Figure 7. Word Cloud Lokawisata Baturaden

4) Word Cloud Musium BRI (The BRI Museum)

Museum BRI is a museum about the first commercial bank in Indonesia. Figure 8 is a word cloud of the BRI Museum. It displays some special keywords such as "sejarah"(history), "bank"(bank), "museum"(museum), "BRI"(the name of one bank in Indonesia). Those keywords inform us that visitors perceived BRI Museum as a historical site.

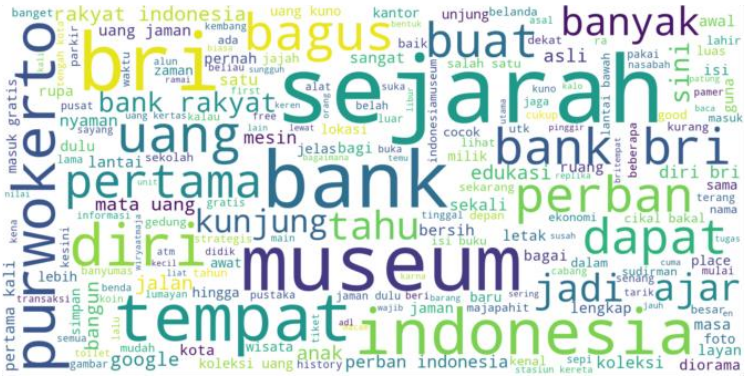

Figure 8. Word cloud Museum BRI

5) Word Cloud Taman Andang (The Andang Park)

Taman Andang is park in countryside of Purwokerto. Based on visitors' comments, the word cloud of Taman Andang is shown in Figure 9. The prominent keywords are "tempat" (place), "banyak"(many), "taman"(park), "buat"(create), "main"(play), "bersih"(clean), "anak"(children). If the noises word such as "taman"(park), "banyak"(many), "buat"(create), and "tempat"(place) are removed, then the main keyword such as clean(bersih), and child(anak) will be more meaningful. We can conclude that visitors consider Taman Andang are clean and appropriate for children.

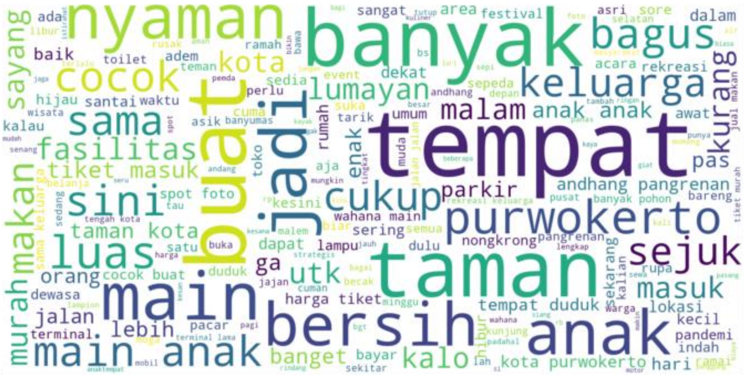

Figure 9. Word Cloud Andang Park (Taman Andang)

6) Word Cloud Curug Telu (The Tree Waterfall)

Curug Telu is natural waterfall in Purwokerto. Tell means three. The Curug Telu word cloud in **Figure 10** displays some unique keywords such as “curug”((waterfalls), “tempat”(place), “bagus”(beautiful), “jalan”(walking), “buat”(create). The word “curug” and “waterfall” are related to the name of this site. Also, "buat"(create) and "tempat"(place) are noise. So, we can neglect all noises. Keyword “bagus”(beautiful) and “jalan”(walking) indicate that visitors consider Curug Telu as a beautiful place and nice to walk.

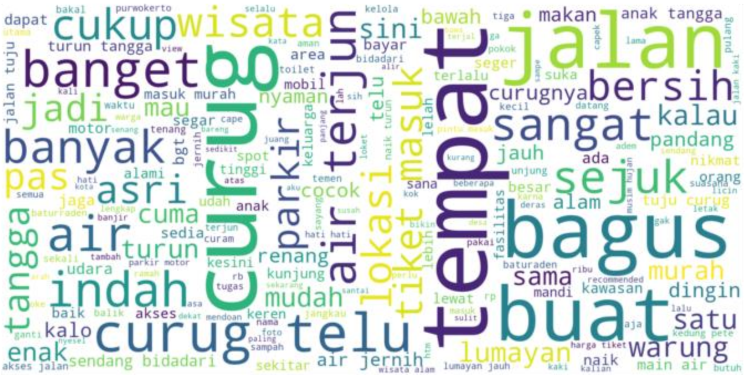

Figure 10. Word Cloud Curug Telu (Tree Waterfall)

7) Word Cloud Telaga Sunyi (The Silence Lake)

Telaga Sunyi is natural lake. Based on data extraction, Telaga Sunyi tourism object has a word cloud that mentions unique words such as "telaga" (lake), "renang"(swimming), "air"(water),"tempat"(place), "bagus"(beautiful), "buat(create), "Alam"(nature),"sepi"(sepi), and "dingin"(cold), shown in **Figure 11.** If the words "sunyi"(silence), "buat"(create), "tempat"(place), "air"(water) are ignored , we gain some meaningful word such as "renang"(swimming), "bagus"(beautiful), "Alam"(nature), and "dingin" (cold). It can see that visitors have an experience at

Telaga Sunyi as a place for swimming, giving a good impression, natural impression, and cold.

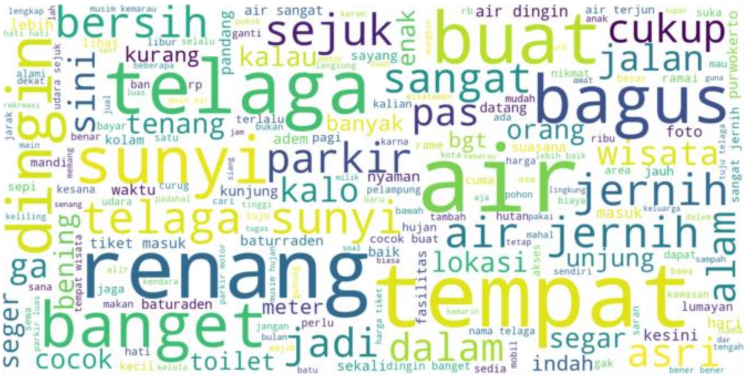

Figure 11. Word Cloud Telaga Sunyi(The Silence Lake)

8) Word Cloud The Village

The Village is an artificial place garden tour. There are many spots like an aviary, an artificial lake, a small zoo, a family restaurant, and a spot of education about livestock. Based on visitors' comments, the word cloud of The Village showed in **Figure 12**. Those word cloud pictured more unique keywords such as “tempat"(place), "bagus"(good), "buat"(make), "foto" (photo), "bayar(payment), "wahana"(rides), "banyak"(many), "makan"(eat). If we ignore the keywords "tempat"(place), "buat"(create), "banyak"(many), We can conclude that visitors have many experiences in The Village. They count that The Village is a nice place to take photographs (selfie, photoshoot), eat, and explore many rides.

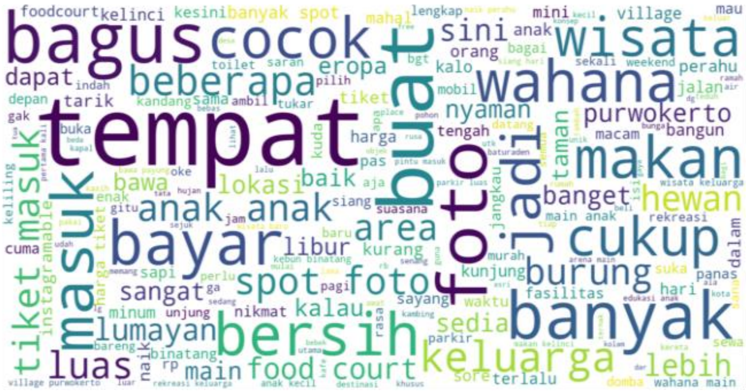

Figure 12. Word Cloud The Village

4. Conclusion

Tourist experiences, especially in eight tourism sites in Purwokerto, have been successfully revealed using sentiment analysis and lexicon-based emotional analysis. Generally, the sentiment of tourists toward all destinations is positive, between 55% and 91%.

Five destinations consider having a high positive sentiment. They are namely the Village (91%), Lokawisata Baturaden(81%), Baturaden Forest (79%) , Limpa Kuwus (78%) and Taman Andang(.77%). At the same time, other destinations achieve positive sentiment under 70%. Only a few visitors give negative sentiment to all tourism destinations. The emotion of visitors stands out in Joy and Trust. On the other hand, NRC revealed sadness and anger even though it is just about 20%.

Word cloud analysis does not reveal a distinguish unique keyword so that the keywords that appeared still contained noise words such as conjunction, adverb, and the name of the sites. Further research must consider making well text preprocessing so that noises words are not processed.

References

[1] A.Gandomi and M.Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

[2] Z.Xiang and D.R.Fesenmaier, "Big Data Analytics , Tourism Design and Smart Tourism," 2017, pp. 299–307.

[3] J.Li, L.Xu, L.Tang, S.Wang, and L. Li, "Big data in tourism research: A literature review," *Tourism Management*, vol. 68, pp. 301–323, 2018, DOI: 10.1016/j.tourman.2018.03.009.

[4] O.Somantri and D.Dairoh, "Analisis Sentimen Penilaian Tempat Tujuan Wisata Kota Tegal Berbasis Text Mining," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 5, no. 2, p. 191, 2019, doi: 10.26418/jp.v5i2.32661.

[5] F.U.Haq and H.Rachmat, "Penggunaan Google Review Sebagai Penilaian Kepuasan Pengunjung Dalam Pariwisata," *Tornare Journal of Sustainable Tourism Research*, vol. 2, no. 1, p. 10, 2020, doi: 10.24198/tornare.v2i1.25826.

[6] S.J.Miah, H.Q.Vu, J. Gammack, and M.McGrath, "A Big Data Analytics Method for Tourist Behaviour Analysis," *Information and Management*, vol. 54, no. 6, pp. 771–785, 2017, DOI: 10.1016/j.im.2016.11.011.

[7] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publisher, 2012.

[8] A.S.Aribowo and S.Khomsah, "Implementation Of Text Mining For Emotion Detection Using The Lexicon Method (Case Study: Tweets About Covid-19)," *Telematika*, vol. 18, no. 1, p. 49, 2021, DOI: 10.31315/telematics.v18i1.4341.

[9] A.Alamsyah and F.N.Zuhri, "Measuring Public Sentiment Towards Services Level in Online Forum using Naive Bayes Classifier and Word Cloud," in *CRS-ForMIND International Conference and Workshop 2017*, 2017, no. October.

[10] M.A.Yudhistira and A.Alamsyah, "Analisis Percakapan Di Media Sosial Twitter Museum Di Indonesia Menggunakan Metode Network Text Analysis Analysis of Conversation in Social Media Twitter of Museum in," in *e{_}Procedding of Management*, 2017, vol. 4, no. 2.

[11] D.F.Budiono, A.S.Nugroho, and A.Doewes, "Twitter sentiment analysis of DKI Jakarta's gubernatorial election 2017 with predictive and descriptive approaches," *Proceedings -*

2017 International Conference on Computer, Control, Informatics, and its Applications: Emerging Trends In Computational Science and Engineering, IC3INA 2017, vol. 2018-Janua, pp. 89–94, 2017, DOI: 10.1109/IC3INA.2017.8251746.

[12] S.Khomsah and Agus Sasmito Aribowo, "Text-Preprocessing Model Youtube Comments in Indonesian," RESTI, vol. 4, no. 4, pp. 648 - 654, Aug. 2020, DOI: 10.29207/resti.v4i4.2035.

[13] S.M.Mohammad,"Sentiment and Emotion Lexicons",https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm (accessedApr.01, 2020).