



Sentiment Analysis on Government Performance in Tourism During The COVID-19 Pandemic Period with Lexicon Based

Adri Priadana¹, Ahmad Ashril Rizal²

¹Center of Data Analytic Research and Services, Universitas Jenderal Achmad Yani
Yogyakarta

²Information Technology, Universitas Islam Negeri Mataram

Email: adripriadana3202@gmail.com, ashril.rizal@gmail.com

ABSTRACT

The COVID-19 pandemic impact has affected all industries in Indonesia and even the world, including the tourism industry. The government has conducted many programs to answer the needs of the tourism industry, especially in making tourism and business destination management programs and carrying out activities oriented, especially during the COVID-19 pandemic. Meanwhile, the government has a role in making policies, especially in the roadmap, for developing the tourism industry. However, the government also needs a way to figure out public sentiments towards the policies that have been implemented. This study aimed to track trending topics and analyze the sentiment of public opinion in Instagram to figure out government performance in tourism during the COVID-19 pandemic period. The results of trending topics will be classified by sentiment analysis using a Lexicon-based and Naive Bayes Classifier. Instagram data taken since January 2020 showed the five highest topics in the tourism sector, namely health protocols, hotels, homes, streets, and beaches. Of the five topics, sentiment analysis was carried out with the Lexicon-based and Naive Bayes classifier, showing that beaches get an incredibly positive sentiment, namely 80.87%, and hotels provide the highest negative sentiment, 57.89%. The accuracy of the Confusion matrix's sentiment results shows that the accuracy, precision, and recall are 82.53%, 86.99%, and 83.43%, respectively.

Keywords: sentiment analysis; government performance in tourism; COVID-19 pandemic period; Lexicon Based

INTRODUCTION

The COVID-19 pandemic impact has affected all industries in Indonesia and even the world, including the tourism industry, and spreads to various other sectors. The tourism industry in Indonesia has links with other sectors such as hotels, restaurants, transportation, and small micro medium enterprises. Furthermore, it impacts on souvenir and culinary entrepreneurs, travel agents, and tour guides. The value of the decline in state revenue in the tourism sector due to COVID-19 on a national scale is, of course, tremendous. The government should not count and study the impact but pay attention to concrete steps in saving the tourism industry in Indonesia.

A careful strategy and planning are needed to save the tourism industry in Indonesia after COVID-19, which can be obtained from social media. One of the social

media that is widely used by Indonesian people is Instagram. From Andi Link¹ shows that Instagram is one of the most-used social media platforms in 2020. In Indonesia, there are 63 million Instagram users in 2020. Data sources from social media are indeed instrumental in research. One of the studies sourced from Instagram data analyzed human selfies by determining several hashtags as a basis. We have shown how image data is detected as a human face using the Haar Cascade method. Image analysis to detect human faces using the Haar Cascade method shows that the applied method produces an accuracy value of 71.48% [1]. Collecting data on Instagram can be done using the web scraping method. Simple additive weighting is successfully applied to the decision support system for selecting endorsement accounts on Instagram. This study's Instagram account parameter include the number of followers, the number of likes, the number of comets, and posts that are always updated. The determination of the best parameters for selecting the Endorse account on Instagram has been successfully carried out, as shown by the system accuracy of 75% [2]. The Instagram platform plays an increasingly central role in social media, essential for users to interact or communicate. Instagram contributes to developing tourism destinations, which it is clear that Instagram and its users are transforming into a new form [3].

Previously, Yadav conducted research related to trip mode's effect on opinions on hotel aspects using a social media analysis approach. Knowing the exact customer expectations will allow service providers to focus more on those aspects that are important. It will help hoteliers prioritize their efforts, allocate resources according to customer needs, and provide tailor-made offers to customers to increase customer satisfaction and optimize resource utilization. With social media being an open source of information, positive or negative sentiments of a hotel's opinion or related aspects can affect a hotel's business. It also provides an opportunity to identify the essential features as perceived by the customer and ascertain what the main reasons for customer dissatisfaction are. The data is taken from hotel visitor reviews based on travel mode on TripAdvisor. TripAdvisor reviews are divided into five modes of travel with very few single travelers (4% of overall reviews), and the majority are family travelers (44%), couples giving reviews 22%, business travelers 19%, and friends 11% [4]. There is another opinion mining platform for extracting and classifying hotel reviews posted by users on tourism websites. The system visits web pages starting at the given URL, extracts reviews from page content, then uses opinion mining to process the content and classifies reviews as positive, negative, and neutral. The proposed process has acceptable accuracy and has the advantage that it does not depend on domains and does not require expensive resources to operate. According to the review, it can be concluded that in the tourism domain, the analysis is made aspect oriented. It is because of the many aspects expressed by users about opinions and mixed sentiments present in a review [5]. The sentiment analysis method is also carried out to research halal tourism in Europe. Halal tourism has recently received significant attention from academics and practitioners. This study analyzes tweets related to halal tourism. This study's findings can help various stakeholders, such as marketers who want to target the halal tourism market. These studies have also contributed to existing knowledge about tourism in general and halal tourism in particular [6].

The government has conducted many programs to answer the needs of the tourism industry, especially in making tourism and business destination management programs and carrying out activities oriented, especially during the COVID-19 pandemic. Meanwhile, the government has a role in making policies, especially in the roadmap, for

¹<https://andi.link/hootsuite-we-are-social-indonesian-digital-report-2020/>

developing the tourism industry. However, the government also needs a way to figure out public sentiments towards the policies that have been implemented. This study aimed to track trending topics and analyze the sentiment of public opinion in Instagram to figure out government performance in tourism during the COVID-19 pandemic period.

Some studies analyzed the sentiment of public opinion in social media that could help the government, or the relevant authorities develop responses or programs to address existing problems in the community, especially during the COVID-19 pandemic. Shofiya and Abidi in 2021 [7] analyzed the sentiment of public opinion about social distancing programs in Canada using Twitter data during the COVID-19 pandemic period. They used the Support Vector Machine (SVM) technique to classify sentiment. Obiedat et al. in 2021 [8] analyzed the sentiment of public opinion to enhance the government decisions in Jordan during COVID-19 Pandemic based on Facebook data. They used Whale Optimization Algorithm & Support Vector Machines (WOA-SVM) methods compared with some other methods to classify sentiment. Habibi et al. in 2021 [9] analyzed the sentiment and modeled the topic of public opinion about COVID-19 epidemics in Indonesia based on Twitter data. They used Latent Dirichlet Allocation (LDA) to model topics and Naive Bayes methods compared with other methods to classify sentiment. Prastyo et al. in 2020 [10] analyzed the sentiment of public opinion on Twitter about the Indonesian Government's Handling of COVID-19. They used SVM with Normalized Poly Kernel to classify sentiment.

In this study, the results of trending topics will be classified by sentiment analysis using a Lexicon-based and Naive Bayes Classifier. There are some studies that use Lexicon-based to analyze sentiment. An Arabic Sentiment Lexicon Built Through Automatic Lexicon Expansion (MoArLex) is a lexicon for a sentiment on a large scale. The lexicon is intended to provide accessible Arabic resources that can be used in sentiment analysis tasks. One of the advantages of using the proposed lexicon and the techniques used in constructing it is that it can include terms commonly used in social media [11]. Meanwhile, Indonesian can use the Sastrawi library, which can be accessed openly. A literature review needs to be carried out to provide information about sentiment analysis studies on social media. The researchers introduced various methods, but the most common methods used in the Lexicon-based method are SentiWordnet and TF-IDF. At the same time, those for machine learning are Naïve Bayes and SVM. Choosing the right sentiment analysis method depends on the data itself [12]. Different preprocessing methods affect the polarity classification of sentiments on Twitter. Removing URLs, deleting stop words, and deleting numbers affects classifier performance minimally. Meanwhile, removing stop words, numbers, and URLs is appropriate for reducing noise but not affecting performance [13].

METHODS

Data Collection

First, we did data extraction to collect Instagram data by using a web data extraction method. It is used to routinely extract data from a web data source [14]. We got the Instagram post data by using hashtags that are near related to tourism in Indonesia. The data is used as a dataset is data with captions based on predetermined hashtags in various languages.

Data Cleaning Process

The second step in this research is the data cleaning process. **Figure 1** shows the data cleaning process, in which there is a processing stage inside of it. In the process, several steps are carried out namely Lowercase, Removing Symbol, Stemming, Tokenizing, and Bag of words.

The lowercase process is the process of converting all the letters in the Instagram caption to lowercase. It functions to facilitate the next process in determining sentiment. **Table 1** shows examples of the lowercase process.

Table 1. Lowercase process

No.	Real Caption	After Lowercase Process
1.	Trust and maximize your vacation with @giliketapang001 guaranteed satisfying fun	trust and maximize your vacation with @giliketapang001 guaranteed satisfying fun
2.	Are you paid for being sent home? What if we open a small business?	are you paid for being sent home? what if we open a small business?
3.	Just bored at home	just bored at home
4.	The holiday has ended. Hopefully, this feeling of happiness can last even though the vacation time is over? The feeling that I want a year off	the holiday has ended. hopefully, this feeling of happiness can last even though the vacation time is over? the feeling that i want a year off

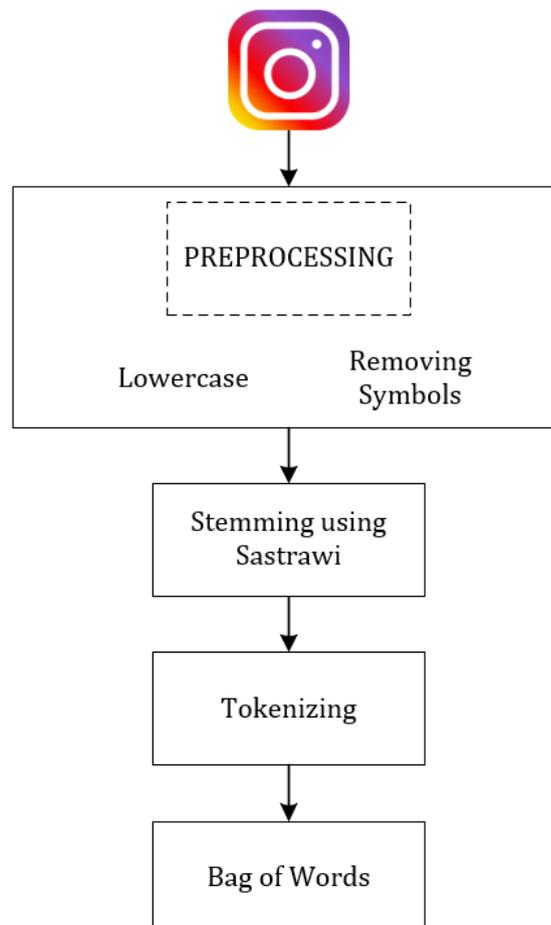


Figure 1. Data cleaning process

The removing symbol process will delete the symbol in the caption. The deleted symbols is a period (.), comma (,), exclamation (!), ask (?), (@), (#), (https: //), and several other unique symbols and emoticons. This process is done because it is assumed that the caption's scraping results do not determine the sentiment results. **Table 2** shows examples of captions after removing symbols.

The stemming process aimed to replace affix words into root words by removing all affixes, whether it's an affix at the beginning, in the middle, or at the end of the word. At this stage, we used the Sastrawi library to replace the affixed word with the root word. Tokenizing process is used to collect the number of words in the data set. The data can be a single word. It means that if there are two words or more than two words in the data set, we only can use one word.

Bag of words is a concept from text analysis. This concept represents the document as an essential information pocket without sorting its words. This method works by counting the total of words' frequency who appeared in a document dataset. So the output of the Bag of words model is a frequency vector.

Table 2. Removing symbol process.

No.	Caption	After Removing Symbol Process
1.	trust and maximize your vacation with @giliketapang001 guaranteed satisfying fun.	trust and maximize your vacation with giliketapang001 guaranteed satisfying fun.
2.	are you paid for being sent home? what if we open a small business?	are you paid for being sent home what if we open a small business
3.	just bored at home.	just bored at home
4.	the holiday has ended. hopefully, this feeling of happiness can last even though the vacation time is over? the feeling that i want a year off.	the holiday has ended. hopefully, this feeling of happiness can last even though the vacation time is over the feeling that i want a year off

Determination of Trending Topics

Trending topics are determined by counting ten captions with the highest frequency. This determination is done by counting the N-grams after getting the word tokenizing results.

Sentiment Analysis using Lexicon Base and Naive Bayes Classifier

In general, the sentiment analysis process is carried out with the steps shown in Figure 2. The Lexicon-based process is carried out after cleaning the previously obtained dataset. The number of words in the data will be calculated based on the rules in the data dictionary. The data dictionary is a collection of words in the Great Dictionary of the Indonesian Language or KBBI and has been classified into negative and positive classes.

Naive Bayes Classifier is a classification with the concept of likelihood when referring to the Bayes hypothesis. Bayes' hypothesis scientifically determines the relationship between the probability of two events A and B, P (A) and P (B) and the likelihood that event A is formed by B and event B is adapted by A, P (A | B) and P (B | A). So the Bayes equation is shown in Equation 1.

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} \tag{1}$$

In this case, the calculated classification is P (A | B), which is the probability that the

hypothesis is correct (valid) for the observed sample B data. The B data is sample data with an unknown class (label), and A is a hypothesis that B is data with a known class (label). P (A) is the probability of hypothesis A, P (B) is the probability of the observed sample data, P (A | B) is the probability of sample data B if it is assumed that the hypothesis is correct (valid) [15].

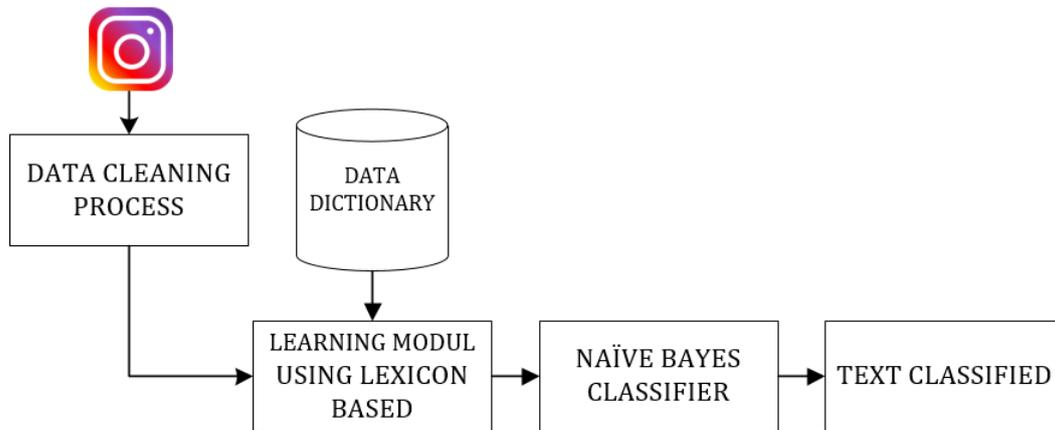


Figure 2. Sentiment analysis proses

Performance Evaluation

In this study, performance evaluation was carried out using the Confusion Matrix as well as precision, recall, and accuracy. Precision and recall values are calculated using Equation 2 and Equation 3. The calculation of predictive accuracy is done using Equation 4 [16],

$$precision = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

$$recall = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (4)$$

RESULTS AND DISCUSSION

This paper's research results are in the form of trending topics in the tourism sector since the COVID-19 pandemic hit. The evaluation results are based on data obtained from Instagram. The data distribution used is shown in Table 3. Data collection was carried out by scraping method using python. The libraries used in natural language processing are the Natural Language Toolkit (NLTK) and the Sastrawi library as cleaning data in Indonesian.

Trending Topics in the Tourism Sector

In this study, the total dataset used is 195136 data. For the first time, the similarity process of each caption is carried out. This function is used because one caption uses a different hashtag, and that hashtag is used as a parameter in this paper. So that one caption will be taken as a dataset if there are the same captions.

The search results for trending topics on Instagram are the accumulated result of Bi-gram and tri-gram using NLTK. There is a word order from each caption that is part of a

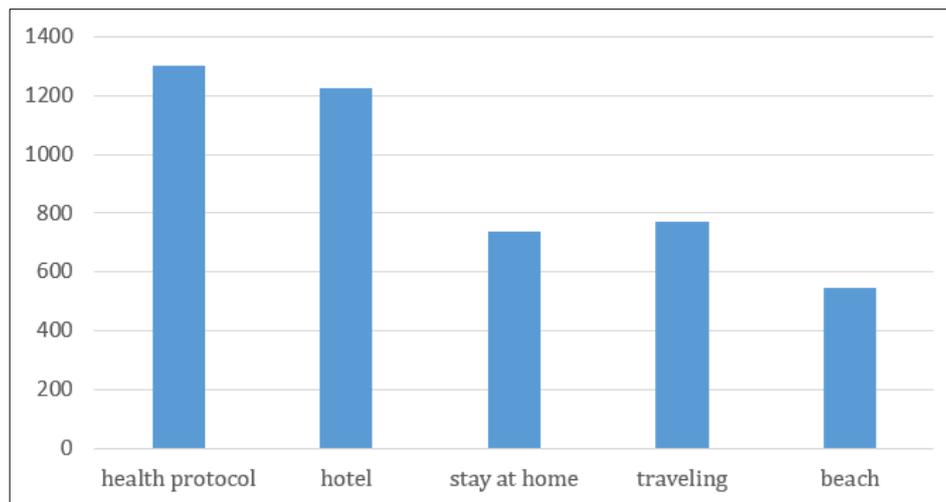


Figure 4. Top 5 Instagram trending topics

Figure 3 and Figure 4 show the results of the trending topic of the data dataset. Figure 3 is a Wordcloud which describes the distribution of text in the dataset. The size of the text is directly proportional to the frequency with which it appears in the dataset. Graphically, Figure 4 shows the top five trending topics based on the dataset used. The trending sequences that emerged were health protocols (1301), hotels (1227), homes (739), roads (772), and beaches (544).

Sentiment Analysis

At this stage, the sentiment results are based on a predefined data dictionary. The Lexion-based method used refers to the data dictionary. In addition to calculating Bayes's probabilities, the number of words in the caption that fall into the negative and positive class is one of the classification references. For example in English translation, "follow the health protocol for the holidays, it seems like it's very complicated" as the caption (1), will be "follow the health protocol for a very complicated holiday" after the cleaning process. In the caption, the words, follow/protocol/health/for/holiday/very are neutral words, and the word /complicated/is negative word.

Figure 5 shows the sentiment results in trending tourism topics on Instagram. Health protocol is the highest topic among Instagram users. However, most wrote positive captions to the health protocol. It is evident from the sentiment calculation results that 73.12% of captions are writing positive things related to health protocols. Not a few Instagram users have complained about health protocols while on vacation or traveling during a pandemic. Still, many Instagram users invite and emphasize complying with health protocols to reduce the number of corona cases in Indonesia. Besides, many hotel complaints not only from tourists but from the hotel and its employees. With the closure of hotel access and the slow pace of hotel openings, many Instagram users have complained about being laid off. Regarding the topic of beaches, many positive captions have appeared. For example, access to the beach is still open even though health protocols are applied. Besides, a caption discusses beaches in Indonesia that are getting prettier and cleaner since the COVID-19 pandemic hit. We can see that beach sentiment is 80.87% positive and only 19.13% with the negative caption.

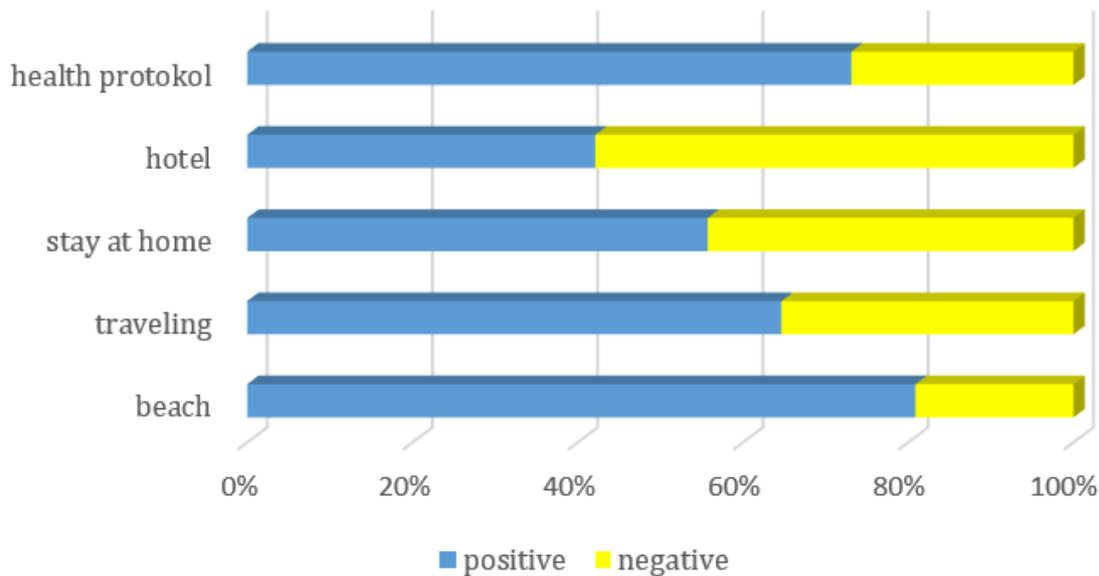


Figure 5. Sentiment analysis of trending topics

Algorithm Performance

The sentiment results in this paper were evaluated using a Confusion Matrix. The results of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are shown in **Table 4**.

Table 4. Confusion matrix of sentiment classification

Class	Classified as Positive	Classified as Negative
Positive	2911	435
Negative	578	1874

$$\text{Precision: } \frac{TP}{TP+FP} * 100\% = \frac{2911}{2911+435} * 100\% = 86.99\%$$

$$\text{Recall: } \frac{TP}{TP+FN} * 100\% = \frac{2911}{2911+578} * 100\% = 83.43\%$$

$$\text{Accuracy: } \frac{TP+TN}{TP+TN+FP+FN} * 100\% = \frac{2911+1874}{2911+1874+435+578} * 100\% = 82.53\%$$

In this study, the lexicon-based method worked by creating a dictionary of opinion words (lexicon) compiled beforehand. The words contained in the data dictionary were divided into two classes, namely classes containing positive words and classes containing negative words. The dictionary was used to identify whether a sentence contains a certain opinion or not. In the text segmented based on word order, the lexicon method would perform a search process on the data dictionary, which words in the text contained positive or negative words.

In the Naive Bayes classifier process, the results of the search for word classes in the data dictionary in the lexicon process would count the number of words in the dataset that fall into the positive class or the words that fall into the negative class. The Naive

Bayes classifier used a prior probability, a true probability value, before carrying out experiments on each label which was the frequency of each label in the training set. The purpose of this process was to classify the caption results on the data in each dataset. The Naive Bayes Classifier method performs the text classification process based on the previously stored training data. In the implementation, there are three stages: making a list of n-grams, then making a list of word classes that have been carried out in the lexicon process and making a classifier. Classifiers need to be trained and to do so requires a list of captions classified into positive and negative classes manually. In its implementation, around 250 positive captions and 250 negative tweet captions were used to train the classifier.

Based on the calculated Confusion Matrix results, we can see that the accuracy of the sentiment results is 82.53%. Precision describes the level of accuracy between the requested data and the predictive results given by the model. The precision result from the sentiment is 86.99%. Meanwhile, the recall which describes the success of the model in recovering considerable information is 83.43%. Some captions should be positive in the negative class or negative captions classified as positive because the Lexicon-based method does not use a learning method. Sometimes in Indonesian captions, there are ambiguous sentences that are positive, but the words chosen to make the caption classified as negative. We can see on an example in the caption (3) of the dataset in English translation like *"for those of you who are nagging about wanting to go on a walk, there are more victims, it's better to stay safe with #dirumahaja"*. Of course, the caption should fall into the positive class. However, words like *nagging* and *victims* are included in the negative data dictionary, and words like *dirumahaja* are classified as positive. Thus, the caption contains more negative words so that it will enter the negative class.

CONCLUSIONS

This study aims to track trending topics in social media Instagram since COVID-19 hit. The results of trending topics will be classified by sentiment analysis using a Lexicon-based and Naive Bayes Classifier. According to analysis results using Instagram data taken since January 2020, it shows the five highest tourism sector topics, namely health protocols, hotels, homes, streets, and beaches. Of the five topics, sentiment analysis was carried out with the Lexicon-based and Naive Bayes classifier, showing that beaches get a very positive sentiment, namely 80.87%, and hotels provide the highest negative sentiment 57.89%. The accuracy of the Confusion matrix's sentiment results shows that the accuracy, precision, and recall are 82.53%, 86.99%, and 83.43%, respectively.

In the future, further investigations will be carried out on how the public perceives the government's performance in the tourism sector during the new normal. Moreover, the lexicon-based method is indeed high-speed in classifying text data. However, the sentiment results are very dependent on the data dictionary used and the caption context. In Indonesia, netizens often write this article with a specific purpose. Captions written using this figure can make the caption fall into the wrong classification. There is a need for a learning method to understand captions with specific figures to fall into the correct classification.

ACKNOWLEDGMENTS

The authors would like to thank the Center of Data Analytic Research and Services, Universitas Jenderal Achmad Yani Yogyakarta, for supporting the data of this research.

REFERENCES

- [1] A. Priadana and M. Habibi, "Face Detection using Haar Cascades to Filter Selfie Face Image on Instagram," in *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIIT)*, 2019, pp. 6–9, doi: 10.1109/ICAIIIT.2019.8834526.
- [2] M. I. Akrianto, A. D. Hartanto, and A. Priadana, "The Best Parameters to Select Instagram Account for Endorsement using Web Scraping," in *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2019, pp. 40–45, doi: 10.1109/ICITISEE48480.2019.9004038.
- [3] M. N. Fatanti and I. W. Suyadnya, "Beyond User Gaze: How Instagram Creates Tourism Destination Brand?," *Procedia - Soc. Behav. Sci.*, vol. 211, pp. 1089–1095, Nov. 2015, doi: 10.1016/j.sbspro.2015.11.145.
- [4] M. L. Yadav and B. Roychoudhury, "Effect of trip mode on opinion about hotel aspects: A social media analysis approach," *Int. J. Hosp. Manag.*, vol. 80, no. September 2018, pp. 155–165, 2019, doi: 10.1016/j.ijhm.2019.02.002.
- [5] C. Bucur, "Using Opinion Mining Techniques in Tourism," *Procedia Econ. Financ.*, vol. 23, no. October 2014, pp. 1666–1673, 2015, doi: 10.1016/s2212-5671(15)00471-2.
- [6] S. Ainin, A. Feizollah, N. B. Anuar, and N. A. Abdullah, "Sentiment analyses of multilingual tweets on halal tourism," *Tour. Manag. Perspect.*, vol. 34, no. January 2019, p. 100658, 2020, doi: 10.1016/j.tmp.2020.100658.
- [7] C. Shofiya and S. Abidi, "Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data," *Int. J. Environ. Res. Public Heal.* 2021, Vol. 18, Page 5993, vol. 18, no. 11, p. 5993, Jun. 2021, doi: 10.3390/IJERPH18115993.
- [8] R. Obiedat, O. Harfoushi, R. Qaddoura, L. Al-Qaisi, and A. M. Al-Zoubi, "An Evolutionary-Based Sentiment Analysis Approach for Enhancing Government Decisions during COVID-19 Pandemic: The Case of Jordan," *Appl. Sci.* 2021, Vol. 11, Page 9080, vol. 11, no. 19, p. 9080, Sep. 2021, doi: 10.3390/APP11199080.
- [9] M. Habibi, A. Priadana, and M. R. Ma'arif, "Sentiment Analysis and Topic Modeling of Indonesian Public Conversation about COVID-19 Epidemics on Twitter," *IJID (International J. Informatics Dev.)*, vol. 10, no. 1, pp. 23–30, Jun. 2021, doi: 10.14421/IJID.2021.2400.
- [10] P. H. Prastyo, A. S. Sumi, A. W. Dian, and A. E. Permanasari, "Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 6, no. 2, pp. 112–122, Oct. 2020, doi: 10.20473/JISEBI.6.2.112-122.
- [11] M. Youssef and S. R. El-Beltagy, "MoArLex: An Arabic Sentiment Lexicon Built Through Automatic Lexicon Expansion," *Procedia Comput. Sci.*, vol. 142, pp. 94–103, 2018, doi: 10.1016/j.procs.2018.10.464.
- [12] Z. Drus and H. Khalid, "Sentiment analysis in social media and its application: Systematic literature review," *Procedia Comput. Sci.*, vol. 161, pp. 707–714, 2019, doi: 10.1016/j.procs.2019.11.174.
- [13] Z. Jianqiang and G. Xiaolin, "Comparison research on text pre-processing methods

- on twitter sentiment analysis," *IEEE Access*, vol. 5, no. c, pp. 2870–2879, 2017, doi: 10.1109/ACCESS.2017.2672677.
- [14] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner, "Web data extraction, applications and techniques: A survey," *Knowledge-Based Syst.*, vol. 70, pp. 301–323, Nov. 2014, doi: 10.1016/j.knosys.2014.07.007.
- [15] S. Rana and A. Singh, "Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques," *Proc. 2016 2nd Int. Conf. Next Gener. Comput. Technol. NGCT 2016*, no. October, pp. 106–111, 2017, doi: 10.1109/NGCT.2016.7877399.
- [16] E. M. Martín and Á. P. del Pobil, *Robust Motion Detection in Real-Life Scenarios*, 1st ed. Springer-Verlag London, 2012.