

Trends in sentiment of Twitter users towards Indonesian tourism: analysis with the k-nearest neighbor method

Eka Purnama Harahap¹, Hindriyanto Dwi Purnomo¹, Ade Iriani¹, Irwan Sembiring¹, Tio Nurtino²

¹Department of Computer Science, Faculty of Informatics Engineering, Satya Wacana Christian University, Salatiga, Indonesia

²Department of Informatics, Faculty of Science and Technology, University of Raharja, Tangerang, Indonesia

Article Info

Article history:

Received Jul 05, 2023

Revised Jan 04, 2024

Accepted Jan 20, 2024

Keywords:

K-nearest neighbor

Sentiment analysis

Twitter

Wonderful Indonesia

ABSTRACT

This research analyzes the sentiment of Twitter users regarding tourism in Indonesia using the keyword "wonderful Indonesia" as the tourism promotion identity. This study aims to gain a deeper understanding of the public sentiment towards "wonderful Indonesia" through social media data analysis. The novelty obtained provides new insights into valuable information about Indonesian tourism for the government and relevant stakeholders in promoting Indonesian tourism and enhancing tourist experiences. The method used is tweet analysis and classification using the K-nearest neighbor (KNN) algorithm to determine the positive, neutral, or negative sentiment of the tweets. The classification results show that the majority of tweets (65.1% out of a total of 14,189 tweets) have a neutral sentiment, indicating that most tweets with the "wonderful Indonesia" tagline are related to advertising or promoting Indonesian tourism. However, the percentage of tweets with positive sentiment (33.8%) is higher than those with negative sentiment (1.1%). This study also achieved training results with an accuracy rate of 98.5%, precision of 97.6%, recall of 98.5%, and F1-score of 98.1%. However, reassessment is needed in the future as Twitter users' sentiments can change along with the development of Indonesian tourism itself.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Eka Purnama Harahap

Faculty of Information Technology, Satya Wacana Christian University

Sidorejo, Salatiga, Jawa Tengah 50711, Indonesia

Email: 982022022@student.uksw.edu

1. INTRODUCTION

Indonesia has immense tourism potential and incredible diversity, making it attractive to both local and international tourists [1], [2]. In an effort to promote Indonesian tourism, the government uses the tagline "wonderful Indonesia" as the identity of Indonesian tourism [3], [4]. In the current digital era, Twitter has become one of the popular platforms for various information and opinions. Based on the We Are Social report in Figure 1, the number of Twitter users in Indonesia reached 18.45 million in 2022 [5]. This number is equivalent to 4.23% of the total Twitter users worldwide, which reached 436 million. These figures place Indonesia as the fifth-largest country in terms of Twitter users globally. Therefore, analyzing the sentiment of Twitter users towards Indonesian tourism using the keyword "wonderful Indonesia" becomes crucial to understand public perspectives on Indonesian tourism.

In this article, the K-nearest neighbor (KNN) method will be used to analyze the sentiment of Twitter users [6]. The aim of this analysis is to determine the level of satisfaction and perspectives of Twitter users towards Indonesian tourism [7]. Research by Syarifuddin [8], regarding public opinion on Twitter on the government's large-scale social restrictions (PSBB) (*pembatasan sosial berskala besar*) (PSBB) policy,

or the policy of restricting movement and stricter community activities which is also commonly known as lockdown in many countries. The algorithm used in the study is also one of the algorithms employed by Syarifuddin [8]. He utilized three algorithms, namely decision tree, KNN, and naïve bayes, with the aim of finding the best accuracy value in the prediction process. Among the three algorithms used, the decision tree algorithm yielded the best results with an accuracy of 83.3%, precision of 79%, and recall of 87.17%.

On the other hand, in a research study conducted on sentiment analysis towards the reopening of tourist destinations amidst the COVID-19 pandemic, the naïve bayes algorithm and the KNN algorithm were utilized to classify tweet data as positive or negative [9]. The research findings revealed that the naïve bayes algorithm achieved the highest accuracy rate of 75.53%, with a positive precision of 71% and a positive recall of 99%. Meanwhile, the KNN algorithm obtained the highest accuracy rate of 48.66%, with a positive precision of 69%, and a positive recall of 69%.

The novelty of this research is a deeper understanding of public sentiment in relation to tourism promotion efforts in Indonesia through the "wonderful Indonesia" campaign, supported by social media data analysis. Along with the development of technology, the important role of social media in shaping public opinions and views is becoming increasingly apparent. Therefore, this article discussed how the KNN method is used in analyzing Twitter user sentiment towards Indonesian tourism using the keyword "wonderful Indonesia" from January 2021 to November 2022. We will delve into the data collection process from Twitter and how the KNN method is implemented in sentiment analysis. Additionally, this research can provide insights into how public feelings and intentions regarding tourism can be influenced by social media and aid in making strategic decisions for the tourism industry [10]. It can also offer information on traveler trends and preferences, thus assisting in tourism product planning and development.

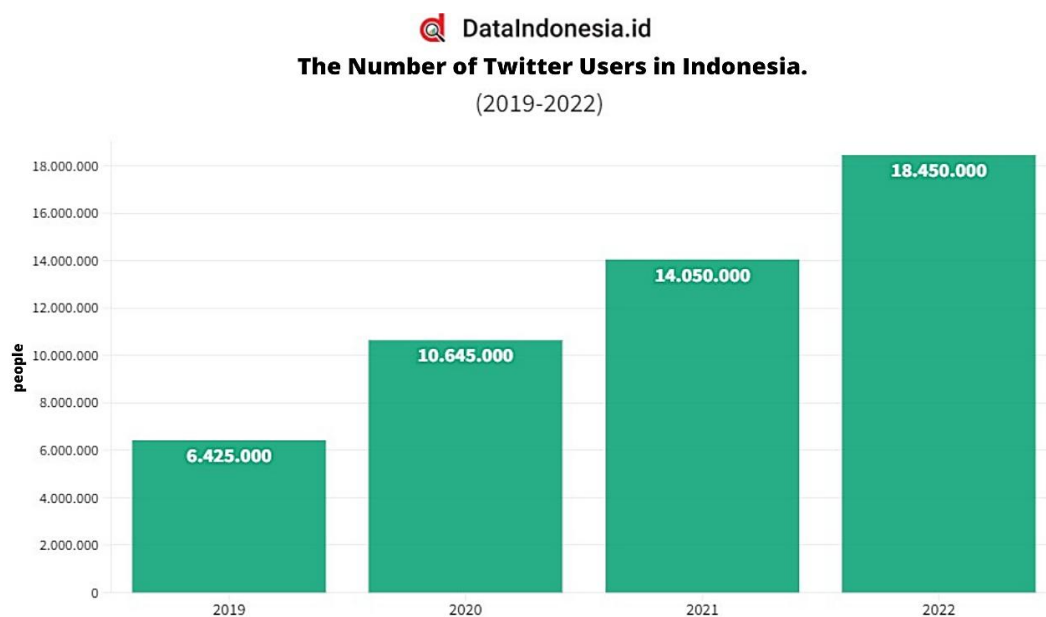


Figure 1. Indonesian Twitter users (2019-2022)

2. METHOD

The method used is the KNN method. As depicted in Figure 2, the KNN method broadly consists of the following steps: scraping, preprocessing, sentiment data, data splitting, training data, testing data, and data visualization [11]. The KNN method was chosen because the stages in system development using the KNN method are considered to be clearly structured.

Apart from the algorithm, the use of larger and more representative datasets is also a key factor in achieving better results. Larger datasets give the model more examples to learn and adapt to a wider variety of language and user communication styles. This allows the model to produce a more accurate and generalized representation of the sentiments expressed in tweets. In addition, new techniques in natural language processing are also applied to derive more informative features from the tweet text, ultimately improving the quality of sentiment prediction.

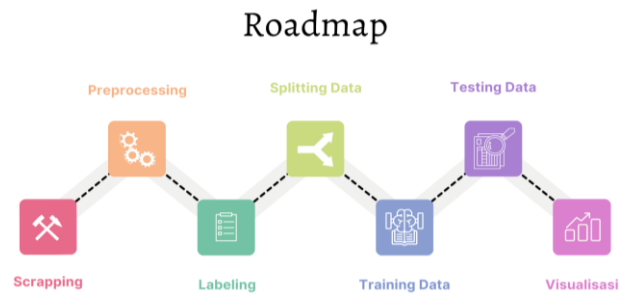


Figure 2. Road map of the KNN method employed

3. RESULTS AND DISCUSSION

3.1. Scraping data

Data scraping is an automation technique used to extract data from websites, databases, enterprise applications, or legacy systems, which can then be saved in a tabular or spreadsheet format [12]. In this study, data scraping was performed on Twitter social media using Google Collaboratory and the Python programming language [13]. The retrieved tweet data was then stored in a CSV file, resulting in a total of 16,543 data. This stage was crucial in obtaining the necessary data for the sentiment analysis of Twitter users towards Indonesian tourism using the keyword “wonderful Indonesia”. Table 1 shows the data successfully collected from the data scraping stage.

Table 1. Data scraping results

Datetime	Tweet	Username
2021-01-30 23:59:29	Last Moments of the Month. #waterfall #Sumaterabarat #wonderfulindonesia #November (In Indonesian)	afrinaldomirfen
2021-01-30 16:19:36	Tourism Greetings! Wonderful Indonesia! Be enthusiastic about taking part, empowered for change! Contact Person: 081213313516 (Mila) Thank you _____ #KabinetDayaJuang #BEMKMUBJ2022 Ministry of Tourism and Creative Economy (HALMAS Coordinating Ministry) BEM KM UBJ 2022-2023 (In Indonesian)	bemkm_ubj
2021-01-30 14:56:10	Hurry up and save it and mark it on your calendar, friend! #WonderfulIndonesia (In Indonesian)	pesonaindonesia
2021-01-30 14:50:38	Love catcher in Bali is fun too. Not bad, Aqua and Wonderful Indonesia are very popular (In Indonesian)	_fleurdevella
2021-01-30 14:49:36	BUMN Minister Erick Thohir: Mutual Cooperation is the Strength of the Indonesian Nation, the Heritage of Our Ancestors. And we must preserve it. . . Wonderful Indonesia...! @erickthohir #RiseTogetherET (In Indonesian)	CryptoKuta

3.2. Preprocessing

Preprocessing is the stage where the obtained tweets are cleaned by removing duplicate tweets, RT, #, @, numbers, emoticons, and other symbols [14]. This stage involves several processes, including data cleansing, tokenization, removal of stopwords, and weighting [15]. The commands used in preprocessing are designed to simplify the application to the data used.

3.2.1. Cleansing

The cleansing stage involves cleaning the data obtained in the sentiment analysis process. The goal is to remove noise and increase the validity of the data before performing sentiment analysis [16]. This stage includes several processes as follows:

- `data = data.drop_duplicates(subset=[''])` to remove duplicate data, which refers to data that have the same sentence. As a result, the obtained data is reduced from 16,543 entries to 14,189 entries.
- `df = data.reset index(drop=True)` to normalize the index order again.
- `Tweet = re.sub('@[A-Za-z0-9]+', '', Tweet)` to remove the mention (@) from the username.
- `Tweet = re.sub('#', '', Tweet)` to remove the hashtag (#).
- `Tweet = re.sub('RT[s]+', '', Tweet)` to remove retweets (RT).
- `Tweet = re.sub('https?:$+', '', Tweet)` to remove the hyperlink (URL).
- `Tweet = re.sub('www.S+', '', Tweet)` to remove the website link.
- `Tweet = re.sub('[A-Za-z ^ ]', '', Tweet)` to remove characters other than the letters A-Z.
- `Tweet = re.sub('(d)', '', Tweet)` to remove digit characters in the Tweet string.
- `Tweet = re.sub('t', '', Tweet)` to remove the tab character in the tweet string.

3.2.2. Tokenizing

Tokenization is an important step in text processing that involves breaking a text or sentence into smaller parts called tokens [17]. Each token is usually the smallest unit of language, such as a word or punctuation mark. The tokenization process aims to break text into elements that are easier for computers to manage and understand. This process is done with the function:

```
from nltk.tokenize import word_tokenize
data['tokens'] = data[text].apply(lambda x: word_tokenize(x))
```

3.2.3. Remove stopwords

Stopwords are a crucial step in text processing that aims to remove words that tend not to contribute significant meaning in a given language [18]. In Indonesian, this process often uses literary libraries to identify and remove words that are considered common and do not provide deep meaning in the context of a sentence or document.

```
def remove_stop_words(tokens):
    filtered_tokens = [token for token in tokens if token not in stop_words]
    return filtered_tokens
def convert_slangword(tokens):
    normalized_words = [slang_dict[word] if word in slang_dict else word for word in tokens]
    return normalized_words
data['stemmed_tokens'] = data['tokens'].apply(remove_stop_words)
data['lemmatized_tokens'] = data['stemmed_tokens'].apply(convert_slangword)
```

3.2.4. Weighting

Weighting is an important process in text analysis that aims to assign weight or importance to words in a document or corpus based on their relative frequency or significance [19]. In text analysis, not all words contribute equally to the meaning or information contained in the text. Therefore, weighting helps distinguish frequently occurring (common) words from infrequently occurring (rare) words, thus creating a more accurate representation of the text. In this stage, the "get word weights" function is used.

3.3. Data labeling

Labeling is the process stage of scanning or labeling the data to identify the sentiment in the data [20]. Labeling is done using the TextBlob library which has an implementation of the sentiment analysis model by calling 'TextBlob(text).sentiment.polarity', it will get a sentiment value between -1 which means negative, 0 which means neutral, and 1 which means positive [21]. At this stage, the sentiment results are obtained with a total of negative 156 data, neutral 9242 data, and positive 4791 data. Table 2 displays the successfully labeled data resulting from the data labeling process. In this stage using functions,

```
def getAnalysis(score):
    if score < 0 :
        return 'Negative'
    elif score == 0 :
        return 'Neutral'
    else:
        return 'Positive'
```

Table 2. Result of data labeling

Datetime	Tweet	Username	Sentiment
2021-01-30 23:59:29	Last Moments of the Month. #waterfall #Sumaterabarat #wonderfulindonesia #November (In Indonesian)	afrinaldomirfen	Neutral
2021-01-30 16:19:36	Tourism Greetings! Wonderful Indonesia! Be enthusiastic about taking part, empowered for change! Contact Person: 081213313516 (Mila) Thank you _____ #KabinetDayaJuang #BEMKMUBJ2022 Ministry of Tourism and Creative Economy (HALMAS Coordinating Ministry) BEM KM UBJ 2022-2023 (In Indonesian)	bemkm_ubj	Positive
2021-01-30 14:56:10	Hurry up and save it and mark it on your calendar, friend! #WonderfulIndonesia	pesonaindonesia	Neutral
2021-01-30 14:50:38	Love catcher in Bali is fun too. Not bad, Aqua and Wonderful Indonesia are very popular (In Indonesian)	_fleurdevella	Positive
2021-01-30 14:49:36	BUMN Minister Erick Thohir: Mutual Cooperation is the Strength of the Indonesian Nation, the Heritage of Our Ancestors. And we must preserve it. . . Wonderful Indonesia...! @erickthohir #RiseTogetherET	CryptoKuta	Neutral

3.4. Splitting data

Splitting data is a process stage to divide the dataset into different parts to perform model evaluation and testing [22]. The splitting process is done by dividing the dataset into two parts, namely the training part and the testing part [23]. Training data is used to train the model and test data is used to evaluate the model. Data splitting is done with a ratio of 8:2, because this process can reduce the risk of overfitting and provide good accuracy. until obtained. This process is done with the help of the scikit-learn library,

```
X = data[['Subjectivity', 'Polarity']]
```

```
y = data['Sentiment']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 0)
```

The data is then split into 10641 train data and 3548 test data.

3.5. Training data

Training data is integral to the formation and development of algorithms in machine learning [24]. The concept involves using a subset of the overall dataset collected to train a model or algorithm so that it can understand the patterns and relationships among the variables. In the context of machine learning, training data is the basis for teaching the model how to make accurate predictions or decisions based on the information at hand. Training is done using the library 'from sklearn.neighbors import KNeighborsClassifier' then using the command,

```
K = 3
```

```
model = KNeighborsClassifier(n_neighbors=K)
```

```
model.fit(X_train, y_train)
```

3.6. Test data

Test data is a portion of the dataset used to evaluate the performance of the model built with the training data [25]. Test data does not participate in the model training process. The test data is accessed using the commands 'x_test' and 'y_test', which contain the test data and the corresponding sentiment labels. The command 'clf.predict(X_test)' is then used to make predictions, and the model's accuracy is evaluated by comparing the predicted results with the actual labels using the command 'accuracy_score(y_test, y_pred)'.

Prediction, in this context, refers to the ability to predict the class or sentiment label of unknown sentiment texts. In sentiment analysis, a model is developed and trained using labeled training data, enabling it to learn patterns and trends that occur in texts with known sentiment. Once trained, the model can be used to predict the sentiment of new, unlabeled texts. After performing the test data, predictions can be made and compared with the actual data using a confusion matrix for better understanding.

```
cm = confusion_matrix(y_test, y_pred)
```

```
class_names = ['Negative', 'Neutral', 'Positive']
```

```
plot_confusion_matrix(model, X_test, y_test, display_labels=class_names)
```

```
plt.xlabel('Predicted')
```

```
plt.ylabel('True')
```

To facilitate readability of the matrix, the sentiment labels -1, 0, and 1 found in model.classes_ are replaced with "negative," "neutral," and "positive" respectively. The results of the above command can be seen in Figure 3.

- Correctly predicted negative label: 0 data
- Negative label predicted as neutral: 9 data
- Negative label predicted as positive: 33 data
- Neutral label predicted as negative: 0 data
- Correctly predicted neutral label: 2279 data
- Neutral label predicted as positive: 19 data
- Positive label predicted as negative: 0 data
- Positive label predicted as neutral: 0 data
- Correctly predicted positive label: 1208 data

Based on the above data, it can be concluded that the most accurate prediction is in the positive label. However, there are still errors in predicting the neutral and negative labels. After that, scoring is calculated using scikit-learn. The resulting calculations are as:

```
Accuracy: 0.9828072153325818
```

```
Precision: 0.9715633303499707
```

```
Recall: 0.9828072153325818
```

```
F1-score: 0.977034183850208
```

The built model shows excellent results in predicting sentiment on the test data. With an accuracy of 98.28%, the model is able to correctly predict almost all of the data used in the testing phase. The precision of 97.16% indicates that the majority of positive predictions made by the model are correct. A recall of

98.28% indicates that the model can accurately identify almost all true positive instances in the dataset. The F1-score of 0.97 demonstrates a good balance between precision and recall. With consistently high performance across all evaluation metrics, it can be concluded that this model is effective in predicting sentiment on the test data.

In comparison with the previously mentioned research results, this study proves excellent quality in predicting sentiment on test data. Research by Syarifuddin [8] in the context of public opinion regarding PSBB policies resulted in the best accuracy of 83.3% with the decision tree algorithm. Research by Era *et al.* [9] related to sentiment towards the reopening of tourist destinations using the naïve bayes and KNN algorithms with naïve bayes accuracy of 75.53% and KNN of 48.66%.

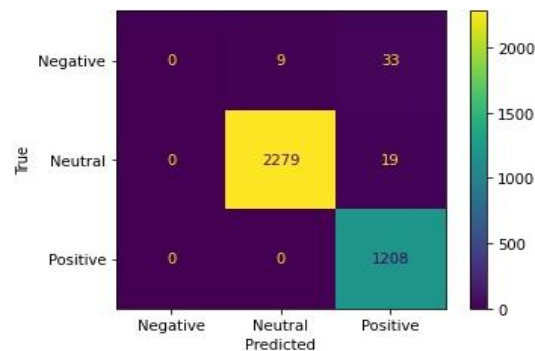


Figure 3. Confusion matrix

3.7. Visualization

Visualization is a key step in data analysis that aims to present information visually through graphs, diagrams, plots, or word cloud. The main goal of visualization is to transform complex data into a more intuitive and understandable representation [26]. Through the use of different types of visualizations, such as bar charts, pie charts, scatter plots, or heat maps, scattered and complex data can be interpreted more clearly and effectively.

3.7.1. Visualization of each month

In Figure 4, the positive sentiment in November has the highest number. It tends to be because the amount of data available in November is more than other months. However, August has more negative sentiment compared to other months and in comparison during January to December the number of negative sentiments is much less than the number of positive sentiments. This shows that the sentiment of Twitter users towards Indonesian tourism has a positive sentiment (a good response to Indonesian tourism).

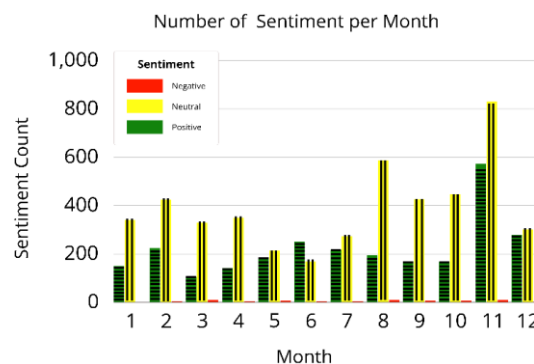


Figure 4. Visualization of the number of sentiments per month

3.7.2. Visualization of sentiment percentage

Visualization of sentiment percentage is done in order to find out what the opinion or sentiment of the Twitter social media tweets of the community is like by looking at the percentage of sentiment obtained.

Figure 5 shows a diagram of the percentage of community sentiment on Twitter social media against the keyword “wonderful Indonesia” The percentage in this chart shows that negative sentiment has the smallest percentage value than positive or neutral sentiment.

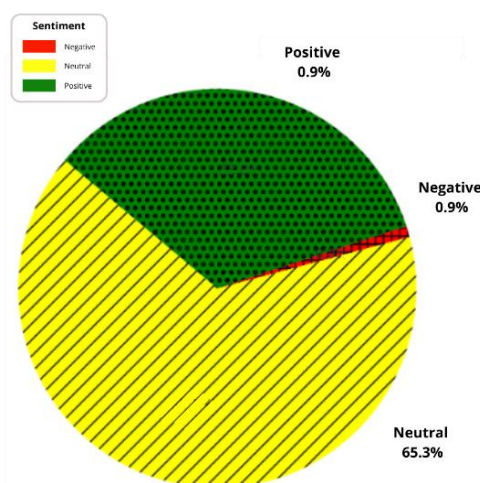


Figure 5. Sentiment percentage from January 2021 to November 2022

3.7.3. Word cloud visualization

Word cloud visualization is a graphical representation of a collection of words, where the size of each word corresponds to its frequency [27]. In this visualization, the most frequently occurring words in the dataset are displayed in larger sizes, while less common words are displayed in smaller sizes. Based on the word cloud in Figure 6, it can be concluded that the words "wonderful," "Indonesia," and "Mandalika" are frequently discussed topics. This is because the data obtained consists mostly of tweets about promoting tourism in Indonesia, which have a neutral sentiment. During that time, there was also a lot of discussion about Mandalika as a new tourist attraction in Indonesia.



Figure 6. WordCloud visualization

3.7.4. Table visualization

Table visualization is a graphical representation of data where the data is organized in columns and rows. Table visualizations allow researchers to clearly display relevant data [28]. The purpose of table visualization is to present data in a structured format. Table 3. shows the results of tweets with negative sentiment.

Table 3. Table tweet sentiment negative

Date	Tweet	Username	Tweet
2021-04-19	Pink Beach, East Lombok, road access is bad, not really, the dirt road is rocky and has holes, making you lazy about planning to go on holiday there (In Indonesian)	BLabuanbajo	Pink Beach, East Lombok, poor road access, not really, rocky dirt roads and potholes, making it lazy to plan a vacation there
	#InIndonesiaAja #WonderfulIndonesia #Lombok #LombokIsland #PulauLombok #WisatadiLombok #TripLombok #LombokTour #PantaiPinik #Pink #PinkBeach https://t.co/LILuM5PSNR		
2021-04-13	Directions on Mount Agung are quite difficult and very prone to misleading climbers when the weather starts to get foggy. #WonderfulIndonesia (In Indonesian)	dombacoklat	The directions on Mount Agung are quite difficult and very prone to misleading climbers when the weather starts to fog up.
2021-03-18	How come there are hotels whose rooms are cramped, damp, have no air ventilation but are still operating and charge high prices, really disappointing (In Indonesian)	indiraysm	How come there are hotels with cramped, humid rooms, no air ventilation but still operating and charging expensive prices, really disappointing.
	#Kemenparekraf #WonderfulIndonesia		
2021-02-26	@ayam_kinantan_reds how are you gbla, there's just a match starting to get a lot of cracks in the walls and the toilets are dirty, especially now it's really bad for sure (In Indonesian)	bubudebo	how are you doing gbla, there is a match just started a lot of cracks in the wall and the toilet is dirty, what's worse now is definitely
	#WonderfulIndonesia #forfootballcultureinindonesia		
2021-02-18	Andi Mattalatta Mattoanging Stadium has the unfortunate fate, abandoned, unused, not surprisingly wonderful Indonesia (In Indonesian)	ayam_kinantan_reds	andi mattalatta matoanging stadium poor fate, abandoned, not repaired, no wonder wonderful indonesia

In the effort to improve management related to negative sentiment towards tourism in Indonesia, there are several suggestions that can be implemented. Firstly, there is a need for infrastructure improvement by repairing poor road access, such as better paving and maintenance. This action will enhance the comfort and safety of tourists when visiting tourist destinations. Furthermore, there is a need for an enhancement in the quality of services in the tourism sector, especially in hotels. Better training for hotel employees, maintenance improvements, and a focus on customer satisfaction will help improve guest experiences and reduce negative sentiment related to poor service.

In addition, efforts are also needed to improve cleanliness and environmental management in tourism. This refers to complaints related to beach cleanliness and waste management. Improving waste management, environmental awareness campaigns, and involving community participation in maintaining cleanliness can help reduce negative sentiment related to environmental cleanliness in tourism. In the context of consumer protection, it is important to enhance transparency in tourism transactions and protect consumers from fraudulent practices or unfair pricing. Clear regulations, strong law enforcement, and education for tourists are necessary to build trust and reduce negative sentiment related to consumer protection.

Lastly, there is a need to improve coordination and communication among stakeholders such as tourism authorities, destination managers, and local communities. Accurate information, clear directions, and open communication will help reduce confusion and enhance the tourists' experience, as well as address complaints related to directions and destination conditions. By implementing these suggestions, it is hoped that tourism management can be enhanced and negative sentiment experienced by tourists can be reduced, thus providing a more positive experience and improving perceptions of tourism in Indonesia.

4. CONCLUSION

Based on the results of the research conducted, it is concluded that the majority of Twitter users' tweets contain neutral sentiments related to tourism in Indonesia. However, if neutral sentiments are ignored, Twitter users' views on tourism tend to be positive. It can be seen that the percentage of positive sentiment reaches 33.8%, while negative sentiment is only 1.1%. This trend is reflected in key words such as "wonderful", "Indonesia", and "Mandalika" that frequently appear in discussion topics. This connection can be seen from the attention given to Mandalika as one of the main tourist destinations in Indonesia in those years. The analysis results using the KNN algorithm showed an accuracy rate of 98.2%, recall 97.1%, precision 98.2%, and F1-score 97.7%, reflecting a highly accurate and valid evaluation for policy purposes. For suggestions in future research, it is recommended to comprehensively consider the following aspects.

One important aspect is to conduct comparisons with other methods and algorithms to test and strengthen the research findings. By adopting a variety of approaches, researchers can gain a broader understanding and increase the reliability of the findings. This approach will allow future research to provide a more comprehensive and in-depth insight into Twitter users' sentiment towards tourism in Indonesia.

ACKNOWLEDGEMENTS

The authors would like to thank the Faculty of Science and Technology, Raharja University and Satya Wacana Christian University Salatiga, who have provided resource persons for the successful implementation of this research. The authors would also like to thank Alphabet Incubator for providing facilities and financial support during this research.

REFERENCES

- [1] A. Minardi, T. Adrian, R. Afriantari, and N. U. Hasanah, "Indonesian Tourism Diplomacy to India," *Indones. J. Tour. Leis.*, vol. 1, no. 1, pp. 1–13, 2020, doi: 10.36256/ijtl.v1i1.83.
- [2] P. Ollivaud and P. Haxton, "Making the most of tourism in Indonesia to promote sustainable regional development," 2019.
- [3] M. Fahrizal, R. Sundari, M. A. A. Shiddiqi, and F. Rani, "The Strategy of the Indonesian Government in Implementing the Wonderful Indonesia Digital Diplomacy Program," *KEMUDI J. Ilmu Pemerintah.*, vol. 7, no. 1, pp. 47–65, 2022.
- [4] U. Salamah and M. R. Yananda, "The Many Faces of Wonderful Indonesia: Tourism Brand in 16 Countries Online News Sites," *J. Komun. Indones.*, vol. 9, no. 1, p. 4, 2020.
- [5] K. Diantoro, A. Soderi, A. Rohman, and A. T. Sitorus, "Sentiment Analysis of Public Opinion on the 2024 Presidential Election in Indonesia Using Twitter Data with the K-NN Method," *Digit. J. Comput. Sci. Appl.*, vol. 1, no. 1, pp. 1–10, 2023.
- [6] M. Wongkar and A. Angdresey, "Sentiment analysis using Naive Bayes Algorithm of the data crawler: Twitter," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, 2019, pp. 1–5.
- [7] Y. Yan, J. Chen, and Z. Wang, "Mining public sentiments and perspectives from geotagged social media data for appraising the post-earthquake recovery of tourism destinations," *Appl. Geogr.*, vol. 123, p. 102306, 2020.
- [8] M. Syarifuddin, "Analysis of public sentiment regarding the effects of PSBB on Twitter using the decision tree, KNN, and Naive Bayes algorithms," (in Indonesian), *INTI Nusa Mandiri*, vol. 15, no. 1, pp. 87–94, 2020.
- [9] D. Era, S. Andryana, and A. Rubhasy, "A Comparison of the Naive Bayes and K-Nearest Neighbor Algorithms in Sentiment Analysis of the Reopening of Tourism During the COVID-19 Pandemic," (in Indonesian), *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 7, no. 1, pp. 263–272, 2023.
- [10] A. Arora, S. Bansal, C. Kandpal, R. Aswani, and Y. Dwivedi, "Measuring social media influencer index-insights from Facebook, Twitter and Instagram," *J. Retail. Consum. Serv.*, vol. 49, pp. 86–101, 2019.
- [11] A. Salma and W. Silfianti, "Sentiment Analysis of User Reviews on COVID-19 Information Applications Using Naive Bayes Classifier, Support Vector Machine, and K-Nearest Neighbor," vol. 6, pp. 158–162, 2021.
- [12] R. Kabacoff, *R in Action: Data Analysis and Graphics with R and Tidyverse*. Simon and Schuster, 2022.
- [13] S. Garg and A. Garg, "Comparison of machine learning algorithms for content based personality resolution of tweets," *Soc. Sci. Humanit. Open*, vol. 4, no. 1, p. 100178, 2021.
- [14] A. L. Hananto, A. P. Nardilasari, A. Fauzi, A. Hananto, B. Priyatna, and A. Y. Rahman, "Best Algorithm in Sentiment Analysis of Presidential Election in Indonesia on Twitter," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 6s, pp. 473–481, 2023.
- [15] G. P. A. Brahmantha and I. W. Santiyasa, "Sentiment Analysis of the Enforcement of PSBB Part II in Jakarta," *JELIKU (Jurnal Elektron. Ilmu Komput. Udayana)*, vol. 9, no. 2, p. 259, 2020, doi: 10.24843/jlk.2020.v09.i02.p13.
- [16] R. Patel and K. Passi, "Sentiment analysis on twitter data of world cup soccer tournament using machine learning," *IoT*, vol. 1, no. 2, p. 14, 2020.
- [17] H. A. Alatabi and A. R. Abbas, "Sentiment analysis in social media using machine learning techniques," *Iraqi J. Sci.*, pp. 193–201, 2020.
- [18] R. Rani and D. K. Lobiyal, "Performance evaluation of text-mining models with Hindi stopwords lists," *J. King Saud Univ. Inf. Sci.*, vol. 34, no. 6, pp. 2771–2786, 2022.
- [19] I. Alsmadi and G. K. Hoon, "Term weighting scheme for short-text classification: Twitter corpuses," *Neural Comput. Appl.*, vol. 31, no. 8, pp. 3819–3831, 2019.
- [20] G. Ansari, C. Saxena, T. Ahmad, and M. N. Doja, "Aspect term extraction using graph-based semi-supervised learning," *Procedia Comput. Sci.*, vol. 167, pp. 2080–2090, 2020.
- [21] K. H. Manguri, R. N. Ramadhan, and P. R. M. Amin, "Twitter sentiment analysis on worldwide COVID-19 outbreaks," *Kurdistan J. Appl. Res.*, pp. 54–65, 2020.
- [22] R. Ahuja, A. Solanki, and A. Nayyar, "Movie recommender system using k-means clustering and k-nearest neighbor," in *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2019, pp. 263–268.
- [23] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. A. Abo-Elsoud, "A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier," *Knowledge-Based Syst.*, vol. 205, p. 106270, 2020.
- [24] L. Wang, "Research and implementation of machine learning classifier based on KNN," in *IOP Conference Series: Materials Science and Engineering*, 2019, vol. 677, no. 5, p. 52038.
- [25] P. Raghavan and N. El Gayar, "Fraud detection using machine learning and deep learning," in *2019 international conference on computational intelligence and knowledge economy (ICCICE)*, 2019, pp. 334–339.
- [26] J. Samuel, G. G. M. N. Ali, M. M. Rahman, E. Esawi, and Y. Samuel, "Covid-19 public sentiment insights and machine learning for tweets classification," *Information*, vol. 11, no. 6, p. 314, 2020.
- [27] T. Turki and S. S. Roy, "Novel hate speech detection using word cloud visualization and ensemble learning coupled with count vectorizer," *Appl. Sci.*, vol. 12, no. 13, p. 6611, 2022.
- [28] T. C. Guetterman, S. Fàbregues, and R. Sakakibara, "Visuals in joint displays to represent integration in mixed methods research: A methodological review," *Methods Psychol.*, vol. 5, p. 100080, 2021.

BIOGRAPHIES OF AUTHORS



Eka Purnama Harahap    is a bachelor graduate from STMIK Raharja, followed by the Information Technology Masters Study Program at STMIK Raharja which since 2019 has changed to University of Raharja. Since 2016 until now, there have been 10+ courses being taught and 2 courses including IT research and introduction to information technology. Hap actively writes scientific papers that are published in national and international reputations. Among them are 50+ accredited national journals indexed by Google Scholar, SINTA, and 5+ SCOPUS indexed international journals. Not only that, she has also been a peer reviewer at 5+ national journals, namely aptisi transactions on technopreneurship and APTIKOM journal on computer science and information technologies (CSIT). She can be contacted at email: 982022022@student.uksw.edu.



Hindriyanto Dwi Purnomo    received his Bachelor of Engineering degree from Gajah Mada University Yogyakarta in physics engineering (intelligent control) in 2005, Master of Information Technology in informatics engineering from the University of Melbourne-Australia in 2009, and Doctor of Philosophy from Chung Yuan Christian University-Taiwan in industrial engineering (numerical optimization) in 2013. His research interests include optimization techniques, machine learning, soft computing, and fuzzy logic. Currently, he is active as a lecturer at the Faculty of Information Technology, Satya Wacana Christian University Salatiga. He can be contacted at email: hindriyanto.purnomo@uksw.edu.



Ade Iriani    from the Satya Wacana Christian University, College Educational Administration Study Program who serves as a lector. She has the status of a permanent lecturer. She can be contacted at email: ade.iriiani@uksw.edu.



Irwan Sembiring    earned a Bachelor of Engineering degree in informatics engineering from the Veteran Yogyakarta National Development University in 2001, a Masters in Computer Science in computer science from Gadjah Mada University Yogyakarta in 2004, and a Doctorate in computer science from Gadjah Mada University Yogyakarta in 2015. His research areas include computer network security and computer network designing. Currently active as a teacher at the Faculty of Information Technology, Satya Wacana Christian University, Salatiga. His main research interests focus on bio-inspired computing, artificial intelligence, metaheuristic modeling and optimization, evolutionary computations, optimization algorithms, information retrieval, feature selection, combinatorial problems, data mining, and text mining. He can be contacted at email: irwan@uksw.edu.



Tio Nurtino    is currently a student pursuing his education in the Informatics Engineering Program at Raharja University. He has a special interest in software development and has actively participated in various related research projects at Raharja University. He aspires to develop innovative solutions that can advance the field of information and communication technology. He can be contacted via email at tio.nurtino@raharja.info.