Scientific Journal of Informatics
Vol. 10, No. 3, Aug 2023

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Coastal Sentiment Review Using Naïve Bayes with Feature Selection Genetic Algorithm

Oman Somantri[1*], Ratih Hafsarah Maharrani[2], Santi Purwaningrum[3]
[1,2]Cybersecurity Engineering Department, Department of Informatics,
State Polytechnic of Cilacap, Indonesia
[3]Informatics Department, Department of Informatics,
State Polytechnic of Cilacap, Indonesia

Abstract.
Purpose: The tourism potential in the maritime sector can be Indonesia's mainstay at this time, especially in enjoying the charm of the natural beauty of the coast as people know Indonesia is an archipelagic country. The purpose of this study is to find the best model by applying the feature selection genetic algorithm (GA) and Information Gain (IG) to get the best Naïve Bayes (NB) model and the best features to produce the best level of sentiment classification accuracy.
Methods: The stages of the research were carried out by going through the process of searching, pre-processing, analyzing research data using the Naïve Bayes model and optimizing genetic algorithms, validating data, and model evaluation.
Results: The experimental results show that the best model is naïve Bayes based on information gain and the genetic algorithm yields an accuracy rate of 86.34%.
Novelty: The main contribution to this research is proposing a new model of the best NB optimization model by applying an optimization algorithm in the search for feature selection to increase sentiment classification accuracy.
Keywords: Coastal, Naïve bayes, Information gain, Feature selection, Genetic algorithm
Received April 2023 / **Revised** April 2023 / **Accepted** May 2023

INTRODUCTION

Indonesia has currently named an archipelagic country that is famous around the world for its biggest potential in the maritime sector. This maritime sector has become one of the prima donnas, both local and foreign tourists deliberately visit to enjoy the natural beauty of Indonesia's coasts. Assessment of people's perspectives on such a beautiful coastal destination will be very influential on the number of people who are interested in following others that have previously seen or enjoyed the beach [1], [2]. Social media is often used by people or tourists in assessing the beauty of beaches [3],[4]. We cannot deny that currently, social media has a big influence compared to other media such as newspapers and other print media [5], [6].

Social media is media that provides an overview of something even if the news is true or not. It is very influential on a person's life and decisions in taking action [7]. Assessment of someone's review of a place is currently often carried out by other people through social media as a means to give a review or assessment of the place therefore others know regardless of whether the review is positive or negative [8]. The same is true with many people reviewing coastal locations, especially in southern coastal areas of Java. The assessment of coastal sentiment reviews that many people do indirectly greatly influences the potential for coastal maritime tourism.

Sentiment analysis is a field of science that utilizes artificial intelligence to enable it to provide decision support in assessing sentiment by categorizing whether the sentiment is positive, negative, or neutral [9]. The application of SA is often carried out using machine learning in which each applied algorithm produces a different level of accuracy according to the strengths and weaknesses of each model. Various fields have used sentiment analysis as a model [10], [11] with the hope that it can help in providing decision support for each policy to be decided. Some examples of research in the field of sentiment analysis research are

[*]Corresponding author.
Email addresses: oman_mantri@yahoo.com (Somantri)
DOI: [10.15294/sji.v10i3.43988](https://doi.org/10.15294/sji.v10i3.43988)

applied to film reviews [12], [13], application reviews on the Google Play Store [14], hotel reviews [15], [16], restaurant customer review [17], and product reviews [18]–[20].

Of the several algorithms often used according to their advantages for sentiment analysis, especially classification, are the neural network algorithm and the naïve Bayes (NB) algorithm. Neural networks have the advantage of being able to carry out learning to work based on the initial experience of the model and can perform calculations in parallel [21], [22]. Naïve Bayes is usually used to apply small-scale data for training, besides has been widely used for data processing, especially text mining since it has a better level of accuracy [23], [24]. Based on its advantages, NB is applied for sentiment reviews of coastal assessments in the hope of producing a good and high level of model accuracy.

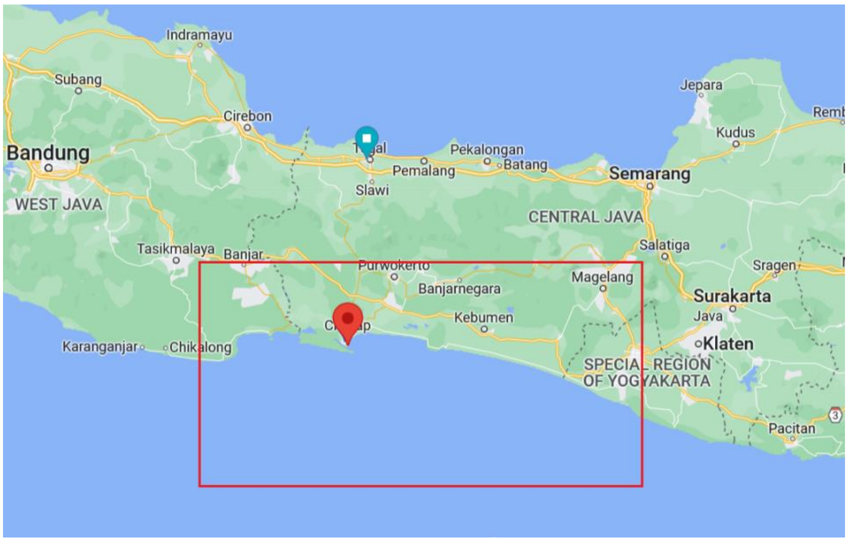

Figure 1. Map of the southern coast of Java Island, Indonesia

The problem that occurs in this algorithm is that there are still parameter values that must be given manually, making it difficult to get the best model and the best weight features. In addition, the determination of the initial weight value in the processed model is still not optimal, so it requires optimization. The purpose of this study is to find the best model by applying an optimization algorithm to get the best model of the two algorithms applied and to get the best weight and parameter values to produce the best level of sentiment classification accuracy.

The research was conducted by Yan, Yingwei, et al [25] using social media as material in assisting planning for the recovery of tourist destinations, especially in the Lombok and Bali areas after the 2018 disaster. The public's view of post-disaster tourist destinations, especially in the Bali and Lombok regions, relies on beach tourism. This study proposes the Latent Dirichlet Allocation (LDA) method with the data from Twitter, and the results of the research show that the proposed and implemented approach can effectively reveal various kinds of sentiments and community perspectives on issues regarding post-disaster tourism recovery from time to time.

Subsequent research was conducted by Park, Eunhye., et al [26] to empirically test the effect of news in predicting the level of tourist arrivals. In this study, data sources originating from news source topics were extracted into data used for forecasting tourist arrivals, especially in Hong Kong. The proposed method is the Autoregressive Integrated Moving Average (ARIMA) method by performing feature selection for selecting variables first. The proposed research model helps tourist destinations in overcoming the externalities of reporting in the news media that affect people's sentiments in assessing a tourist destination. Another study related to destination sentiment was conducted by Ali, and Twiland [27] to get the best model of tourist experience sentiment in Morocco. In this study, a combined model is proposed using a combination of topic modeling and lexicon-based algorithms using Latent Dirichlet Allocation (LDA) where the data comes from TripAdvisor reviews of various tourist attractions in Marrakech, Morocco. The next research is slightly different, which was conducted by Sohrabi, B [28] that proposed a model for predicting tourist destination visits based on comments and interests using text mining by the X-means clustering model and classification with a Decision Tree.

Based on some of the results of previous studies, the researchers conducted experiments using an algorithmic model without optimizing the model, therefore the resulting level of accuracy was not optimal. On the other hand, these limitations are accompanied by the source of the research dataset used in the form of a review of tourist attractions. It use social media data and its influence has an impact on the pre-processing of data that leads to different resulting level of accuracy. For this reason, one of the efforts to increase the level of accuracy produced by the sentiment review model of the coast as a tourist attraction is to propose a new model using Naïve Bayes (NB). This article proposes Naïve Bayes for the classification of sentiment review destinations on the southern coast of Java Island as a recommendation to increase maritime tourism visits, especially beach tourism based on feature selection using a genetic algorithm. The main contribution to this study is to apply an optimization algorithm for feature selection in the NB model to increase the accuracy of the sentiment review classification.

METHODS

Dataset

The research carried out is using experimental research methods, where the data will be processed and input into the model to get the model with the best level of accuracy. The data used in this study were taken from the website *https://www.google.com/maps*, by entering the keyword "beach" which then the search results contained a review of places according to the desired beach name with different star rating values. The difference is between 1 to 5 where the rating value of 5 is the highest positive sentiment. The data taken is Indonesian language text data taken from 2018 to 2021 which is then processed into the desired model of 390 data, examples of data taken are shown in Table 1.

Table 1. Example of a research dataset used in Bahasa

Sentiment Positive	Sentiment Negative
• [Pantai teluk penyu Cilacap, tempat yang bagus untuk melihat luasnya lautan dan mendengarkan deburan ombak laut. Datang ke tempat ini pas bukan waktu liburan jadi lumayan agak lenggang, banyak penjual kuliner disekitar tempat ini, juga ada beberapa tempat untuk berteduh para pengunjung dari teriknya matahari.]	• [Pantainya jorok, tidak ada tempat sampah, masyarakat dan pengelola pantai harus sadar bebersih.]
• [Pantai yg ramah buat ciblonan.. walau agak kotor tapi masih tetap indah. Kita juga bisa menyebrang naik perahu menuju pantai pasir putih, banyak batu2 kecil yg indah berwarna warni.]	• [Pantai nya kotor banget, beli mendoan seporsi isi 8 biji 28.000, wkwk padahal di pantai sodong mendoan seporsi isi 8 cuma 15.000.]

The process of determining to label in this study was carried out based on the assessment of the asterisk. It is included in negative sentiment if the data is given 1-3 stars, and it is included in the positive sentiment category if the data is rated 5 stars. In this study, the sentiment sought in the model is only limited to positive and negative sentiments.

Preprocessing Data

In the proposed research, the next step after the dataset in the form of a coastal review text is obtained is to do data preprocessing. The stages of the preprocessing process in this study were carried out to obtain the expected text data, namely data cleansing, tokenization, stemming, and data filtering [29], [30]. The next process is the data is input into a predetermined model, namely the model using Naïve Bayes.

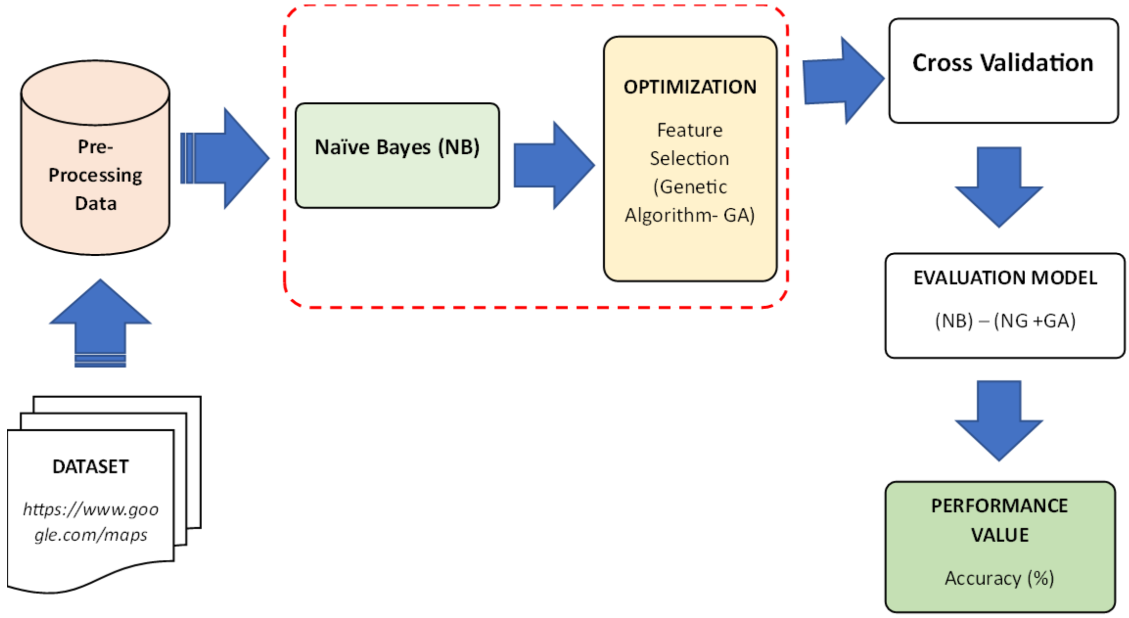

Figure 2. The proposed method framework

The next stage is to carry out model optimization using the feature selection algorithm and conduct data validation using the cross-validation method. It is expected that there will be an increase in the accuracy of sentiment classification and it will get better [31]. The process of the stages carried out in this study can be seen in Figure 2.

Naïve Bayes Algorithm

Naïve Bayes is a method that is included in a classification algorithm derived from the concept of processing statistical data and probabilities that can predict these opportunities based on previous experience [32], [33]. The NB equation is shown in equation (1) [34].

$$P(H|X) = \frac{P(X|H), P(H)}{P(X)} \tag{1}$$

In equation (1) where X is data with unknown class, H is hypothesis data X, P(H|X) is the probability of hypothesis H based on X conditions, P(H) is the hypothesis probability of H, P(X |H) is the probability of X based on the conditions in the hypothesis H, and P(X) is the probability of X.

To get a performance value from the sentiment classification obtained, this study uses equation (2) [35], [36].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

In equation (2), where TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

RESULTS AND DISCUSSIONS

Experimental Results Using Naïve Bayes

The first stage is to experiment with several methods using a predetermined algorithm that is proposed to obtain an expected model. The Naïve Bayes algorithm is applied to the preliminary experiment, in this process NB is expected to give the best accuracy rate of the proposed model. The results of applying the NB model to the experiment are shown in Table 2.

Table 2. Experimental Results Using Naïve Bayes

Fold	Accuracy		
	Stratified Sampling (%)	Shuffled Sampling (%)	Linear Sampling (%)
10	61.79	62.31	55.90
9	63.34	61.83	55.41
8	63.87	62.08	54.11
7	64.10	62.80	54.61
6	**67.44**	61.54	53.08
5	62.05	64.10	51.03
4	63.57	64.62	51.35
3	67.44	63.33	50.26
2	65.13	63.85	42.56

As shown in Table 2, it can be seen that the highest level of accuracy obtained was 67.44% which was obtained using the Fold=6 parameter with the stratified sampling method. In this experiment, the results still need improvements therefore at a later stage another model was implemented for optimization.

Naïve Bayes Experiment Based on Information Gain (IG)

The next stage of the process is to optimize the model to increase the accuracy of the coastal review sentiment classification. At this stage, an optimization process is carried out using the Information Gain (IG) method [37], [38], so that an increase in accuracy is obtained and the results can be seen in Table 2 and Figure 2.

Based on Figure 3 and Table 3, it can be seen that there is a significant change. The lowest accuracy level produced is 58.72% by using fold = 2 and linear sampling. The highest level of accuracy in this model changes quite high, this change is an effort that has been made, namely by applying Information Gain to the Naïve Bayes algorithm so that there is an increase in the accuracy level of 80.91%. This change occurs by using the parameter fold=9 and using linear sampling.

Table 3. Naive Bayes model experiment and information gain

Fold	Accuracy		
	Stratified Sampling (%)	Shuffled Sampling (%)	Linear Sampling (%)
10	73.08	72.05	79.49
9	74.88	71.49	**80.91**
8	75.89	71.82	79.76
7	70.54	75.16	75.10
6	71.54	74.36	76.92
5	72.31	70.26	73.33
4	64.84	75.89	61.36
3	74.87	69.49	66.15
2	64.62	67.18	58.72

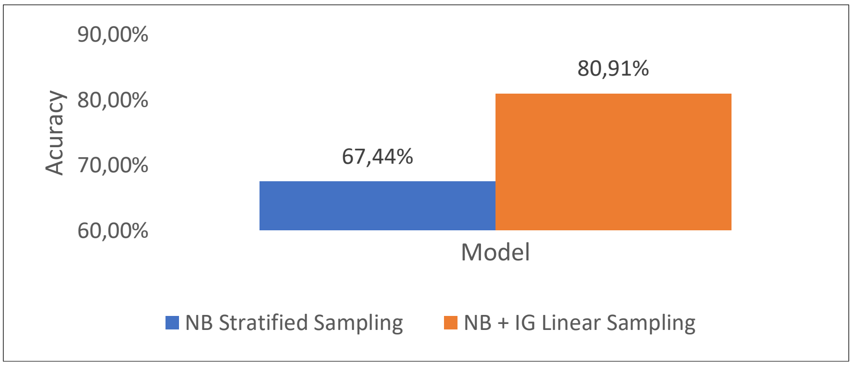

Figure 3. Comparison of the application of the NB and NB+IG methods

Optimization of Feature Selection

Efforts are being made to obtain a coastal sentiment review classification model. One other effort is to apply an algorithm for feature selection, namely by using the genetic algorithm (GA) method. The genetic algorithm is one of the optimization algorithms with its ability to be used for classification. It is used as a feature selection to select the best features in the NB model so that the resulting level of accuracy is better [39].

Based on the experimental results, a significant result was obtained compared to the previous one, shown in Table 4 to produce the best model by determining the parameter values in the GA algorithm, namely population = 5 and selection scheme = tournament.

Table 4. Experimental results NB+IG and GA, tournament.

Crossover type	Fold	Sampling	Accuracy
uniform	9	linear	**86.34%**
uniform	10	linear	81.54%
uniform	9	shuffled	77.13%
uniform	10	shuffled	76.92%
uniform	9	stratified	77.17%
uniform	10	stratified	77.18%
shuffled	9	linear	85.29%
Shuffled	10	linear	85.64%
Shuffled	9	shuffled	76.18%
Shuffled	10	shuffled	75.90%
Shuffled	9	stratified	76.17%
Shuffled	10	stratified	75.90%

Table 5. Experimental results NB+IG and GA, roulette wheel

crossover type	Fold	Sampling	Accuracy
uniform	9	linear	80.78%
uniform	10	linear	83.33%
uniform	9	shuffled	75.88%
uniform	10	shuffled	76.15%
uniform	9	Stratified	76.14%
uniform	10	Stratified	75.64%
shuffled	9	linear	76.11%
shuffled	10	linear	**85.64%**
shuffled	9	shuffled	75.92%
shuffled	10	shuffled	76.67%
shuffled	9	Stratified	76.15%
shuffled	10	Stratified	76.41%

To compare feature selection optimization, another experiment was carried out using different parameters, namely population = 5 and selection scheme = roulette wheel which produced experimental results as shown in Table 5.

The best model for optimizing the NB method using a genetic algorithm-based feature selection is 86.34% with a micro average accuracy of 86.41%. This increase in accuracy has a significant impact on model accuracy, starting from 80.91% to an increase of 5.43%. The experimental results obtained are shown in Table 6, besides that by using formula (2) the micro average accuracy is 86.41%.

Table 6. Confusion matrix result

	True Negative	True Positive	Class Precision
Prediction negative	100	13	88,50%
Prediction Positive	40	237	85,56%

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Accuracy = \frac{237 + 100}{237 + 100 + 40 + 13}$$

$$Accuracy = \frac{337}{390} = 0{,}8641$$

To evaluate the performance evaluation of the model obtained, the AUC (Area Under the Curve) value is applied [40]. In this best model, it was found that the AUC value was 0.79 and the AUC value for this model is included in the "fair classification" criteria in the table. The AUC category table itself is shown in Table 7.

Table 7. AUC Value, its meaning, and symbols

AUC Range	Classification level
0,9 – 1.0	Excellent classification
0,8 – 0,9	Good classification
0,7 – 0.8	Fair classification
0,6 – 0,7	Poor classification
0,5 – 0,6	Failure

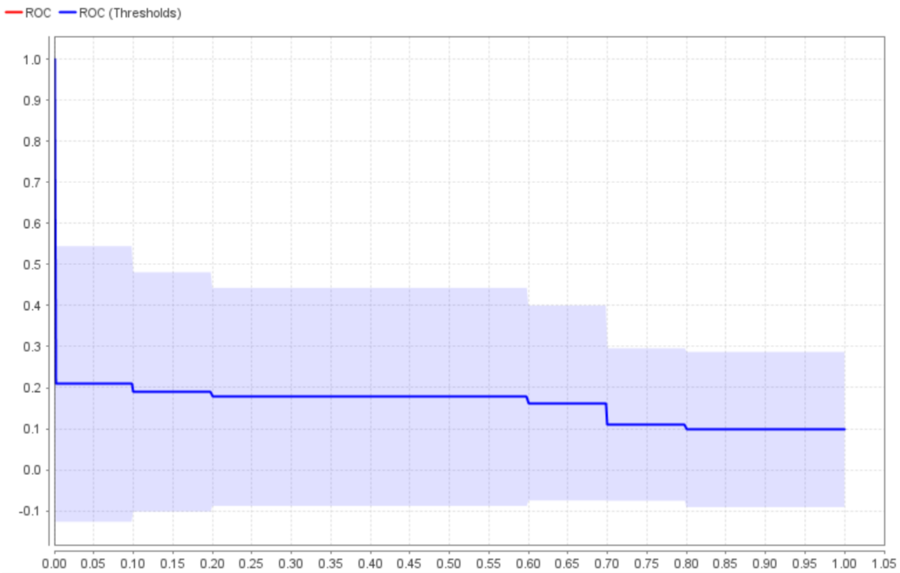

Figure 4. AUC value of the proposed model

Although the model proposed and obtained using feature selection-based Naïve Bayes is currently in the "fair" category, in terms of accuracy it still produces a fairly high value of 86.35%. However, in the future, it still requires an increase in accuracy, especially the resulting AUC value.

Evaluation Model

The results of the experiments that have been carried out show different levels of accuracy. This provides evidence that the level of accuracy obtained does not depend on just one model or algorithm that can be applied, but still requires optimization efforts that can be maximized if the desired model is not sufficient or in accordance. Changes in the parameter values in each model greatly influence the level of accuracy and this makes a great effort in determining the parameter values to match what we expect.

Based on the experimental results, the evaluation of the model was carried out by comparing several experimental methods that obtained the results obtained using classic Naïve Bayes (NB), NB with Information Gain, and NB and Information Gain optimized using a genetic algorithm (GA). The results of the model evaluation are shown in Figure 5 and Table 8.

Table 8. Evaluate the naïve Bayes model

No.	Model	Akurasi
1	Naïve Bayes	67,44%
2	Naïve Bayes + IG	80,91%
3	NB + IG + GA - tournament	**86,34%**
4	NB + IG + GA - roulete wheel	85,64%

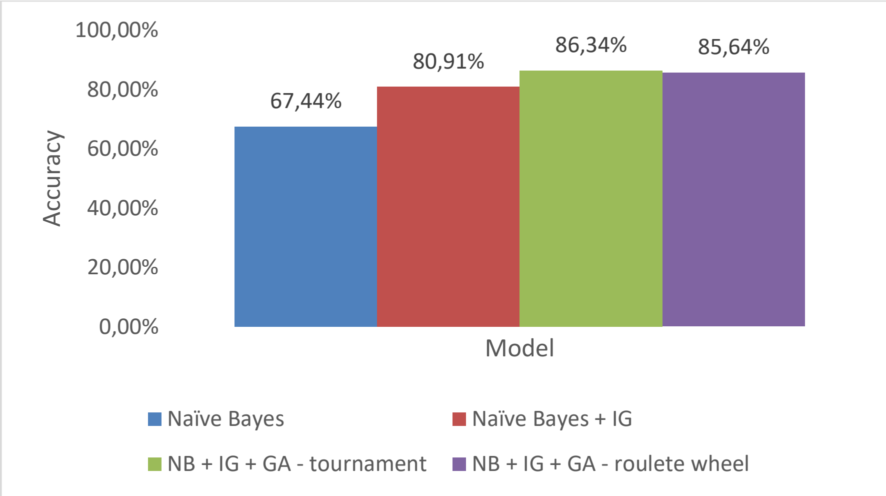

Figure 5. Coastal review sentiment model evaluation chart

Based on Table 6, it can be seen that if we evaluate and compare the several models that have been obtained, it can be seen that the highest accuracy rate is 86.34% using the NB_IG algorithm based on feature selection using GA using selection scheme = tournament, and the folds used are 9, and population = 5. Based on the results obtained in this case, the proposed model is a model that has a greater degree of accuracy compared to other models.

CONCLUSION

The sentiment review assessment model for the coast using the Naïve Bayes algorithm has been obtained after optimization with the highest accuracy rate of 86.34%. The model obtained provides a benefit that can be used by policymakers, especially related parties, to improve coastal maritime tourism, be it services, facilities, or other things that can be optimized. The level of accuracy produced at this time requires efforts to improve accuracy, so further research are needed. It is recommended for further research to optimize from various angles such as data pre-processing, selectinging the best parameters, and optimizing weight values. In addition, it is necessary to apply other algorithms to obtain other experimental results therefore the best model can be seen and applied.

ACKNOWLEDGEMENT

Thank you to the Ministry of Education and Culture through the Academic Directorate of Vocational Education, and the Directorate General of Vocational Education, for funding this research through the Beginner Lecturer Research scheme for the 2022 implementation year.

REFERENCES

[1] V. Teles da Mota, C. Pickering, and A. Chauvenet, “Popularity of Australian beaches: Insights from social media images for coastal management,” *Ocean Coast. Manag.*, vol. 217, p. 106018, Feb. 2022, doi: 10.1016/j.ocecoaman.2021.106018.

[2] I. I. Ibrahim, “Social Media Analysis in Building Customer Trust - Systematic Literature Review,” in *Proceedings of 2022 International Conference on Information Management and Technology, ICIMTech 2022*, Aug. 2022, pp. 39–44, doi: 10.1109/ICIMTech55957.2022.9915099.

[3] M. T. Cuomo, I. Colosimo, L. R. Celsi, R. Ferulano, G. Festa, and M. La Rocca, “Enhacing

Traveller Experience in Mobility Services Via Big Social Data Analytics," *Technol. Forecast. Soc. Change*, vol. 176, p. 121460, Mar. 2022, doi: 10.1016/j.techfore.2021.121460.

[4] N. Pruksorranan, "Conceptual Perspectives on Tourist Behavior toward Information Technology and Social Media in Bangsaen Beach, Chonburi, Thailand," in *2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, Mar. 2020, pp. 58–60, doi: 10.1109/ECTIDAMTNCON48261.2020.9090748.

[5] F. Mirzaalian and E. Halpenny, "Exploring destination loyalty: Application of social media analytics in a nature-based tourism setting," *J. Destin. Mark. Manag.*, vol. 20, p. 100598, Jun. 2021, doi: 10.1016/j.jdmm.2021.100598.

[6] D. Obembe, O. Kolade, F. Obembe, A. Owoseni, and O. Mafimisebi, "Covid-19 and the tourism industry: An early stage sentiment analysis of the impact of social media and stakeholder communication," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 2, p. 100040, Nov. 2021, doi: 10.1016/j.jjimei.2021.100040.

[7] R. Dolan, Y. Seo, and J. Kemper, "Complaining practices on social media in tourism: A value co-creation and co-destruction perspective," *Tour. Manag.*, vol. 73, pp. 35–45, Aug. 2019, doi: 10.1016/j.tourman.2019.01.017.

[8] W. He, G. Yan, J. Shen, and X. Tian, "Developing a workflow approach for mining online social media data," in *2017 IEEE SmartWorld Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017 - Conference Proceedings*, Aug. 2018, pp. 1–6, doi: 10.1109/UIC-ATC.2017.8397648.

[9] M. B. Nasreen Taj and G. S. Girisha, "Insights of strength and weakness of evolving methodologies of sentiment analysis," *Glob. Transitions Proc.*, vol. 2, no. 2, pp. 157–162, Nov. 2021, doi: 10.1016/j.gltp.2021.08.059.

[10] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers," *Computer Science Review*, vol. 27. pp. 16–32, 2018, doi: 10.1016/j.cosrev.2017.10.002.

[11] D. M. E. D. M. Hussein, "A survey on sentiment analysis challenges," *J. King Saud Univ. - Eng. Sci.*, vol. 30, no. 4, pp. 330–338, 2018, doi: 10.1016/j.jksues.2016.04.002.

[12] U. I. Larasati, M. A. Muslim, R. Arifudin, and A. Alamsyah, "Improve the Accuracy of Support Vector Machine Using Chi Square Statistic and Term Frequency Inverse Document Frequency on Movie Review Sentiment Analysis," *Sci. J. Informatics*, vol. 6, no. 1, pp. 138–149, May 2019, doi: 10.15294/sji.v6i1.14244.

[13] Y. Gao, M. Gong, Y. Xie, and A. K. Qin, "An Attention-Based Unsupervised Adversarial Model for Movie Review Spam Detection," *IEEE Trans. Multimed.*, vol. 23, pp. 784–796, 2021, doi: 10.1109/TMM.2020.2990085.

[14] S. Fransiska, R. Rianto, and A. I. Gufroni, "Sentiment Analysis Provider By.U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method," *Sci. J. Informatics*, vol. 7, no. 2, pp. 203–212, Nov. 2020, doi: 10.15294/SJI.V7I2.25596.

[15] N. P. Ririanti and A. Purwinarko, "Implementation of Support Vector Machine Algorithm with Correlation-Based Feature Selection and Term Frequency Inverse Document Frequency for Sentiment Analysis Review Hotel," *Sci. J. Informatics*, vol. 8, no. 2, pp. 297–303, Nov. 2021, doi: 10.15294/sji.v8i2.29992.

[16] D. Apriliani, T. Abidin, E. Sutanta, A. Hamzah, and O. Somantri, "SentiHotel: a sentiment analysis application of hotel services using an optimized neural network," *Bull. Electr. Eng. Informatics*, vol. 10, no. 3, Jun. 2021, doi: 10.11591/eei.v10i3.3040.

[17] O. Somantri and D. Apriliani, "Opinion mining on culinary food customer satisfaction using naïve bayes based-on hybrid feature selection," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 15, no. 1, pp. 468–475, Jul. 2019, doi: 10.11591/ijeecs.v15.i1.pp468-475.

[18] H. T. Ismet, T. Mustaqim, and D. Purwitasari, "Aspect Based Sentiment Analysis of Product Review Using Memory Network," *Sci. J. Informatics*, vol. 9, no. 1, pp. 73–83, May 2022, doi: 10.15294/sji.v9i1.34094.

[19] Y. Wang and X. Li, "Mining Product Reviews for Needs-Based Product Configurator Design: A Transfer Learning-Based Approach," *IEEE Trans. Ind. Informatics*, vol. 17, no. 9, pp. 6192–6199, Sep. 2021, doi: 10.1109/TII.2020.3043315.

[20] Z. Zhao, J. Wang, H. Sun, Y. Liu, Z. Fan, and F. Xuan, "What Factors Influence Online Product

Sales? Online Reviews, Review System Curation, Online Promotional Marketing and Seller Guarantees Analysis," *IEEE Access*, vol. 8, pp. 3920–3931, 2020, doi: 10.1109/ACCESS.2019.2963047.
[21] R. K. Behera, M. Jena, S. K. Rath, and S. Misra, "Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data," *Inf. Process. Manag.*, vol. 58, no. 1, p. 102435, Jan. 2021, doi: 10.1016/j.ipm.2020.102435.
[22] W. Liao, B. Zeng, J. Liu, P. Wei, X. Cheng, and W. Zhang, "Multi-level graph neural network for text sentiment analysis," *Comput. Electr. Eng.*, vol. 92, p. 107096, Jun. 2021, doi: 10.1016/j.compeleceng.2021.107096.
[23] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," *Futur. Gener. Comput. Syst.*, vol. 106, pp. 92–104, May 2020, doi: 10.1016/j.future.2020.01.005.
[24] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, "Variable selection for Naïve Bayes classification," *Comput. Oper. Res.*, vol. 135, p. 105456, Nov. 2021, doi: 10.1016/j.cor.2021.105456.
[25] 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, Oct. 2020, doi: 10.1016/j.apgeog.2020.102306.
[26] E. Park, J. Park, and M. Hu, "Tourism demand forecasting with online news data mining," *Ann. Tour. Res.*, vol. 90, p. 103273, Sep. 2021, doi: 10.1016/j.annals.2021.103273.
[27] T. Ali, B. Marc, B. Omar, K. Soulaimane, and S. Larbi, "Exploring destination's negative e-reputation using aspect based sentiment analysis approach: Case of Marrakech destination on TripAdvisor," *Tour. Manag. Perspect.*, vol. 40, p. 100892, Oct. 2021, doi: 10.1016/j.tmp.2021.100892.
[28] B. Sohrabi, I. Raeesi Vanani, N. Nasiri, and A. Ghasemi Rud, "A predictive model of tourist destinations based on tourists' comments and interests using text analytics," *Tour. Manag. Perspect.*, vol. 35, p. 100710, Jul. 2020, doi: 10.1016/j.tmp.2020.100710.
[29] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Inf. Fusion*, vol. 36, pp. 10–25, 2017, doi: 10.1016/j.inffus.2016.10.004.
[30] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Inf. Process. Manag.*, vol. 50, no. 1, pp. 104–112, Jan. 2014, doi: 10.1016/j.ipm.2013.08.006.
[31] J. Camacho and A. Ferrer, "Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: Practical aspects," *Chemom. Intell. Lab. Syst.*, vol. 131, pp. 37–50, 2014, doi: 10.1016/j.chemolab.2013.12.003.
[32] P. Cichosz, "Naïve Bayes classifier," in *Data Mining Algorithms*, Chichester, UK: John Wiley & Sons, Ltd, 2015, pp. 118–133.
[33] S. Sona. D, S. Asha D, and B. Samarjeet, "5 Assimilate Machine Learning Algorithms in Big Data Analytics: Review," in *Applications of Machine Learning in Big-Data Analytics and Cloud Computing*, River Publishers, 2021, pp. 81–114.
[34] S. Ruan, H. Li, C. Li, and K. Song, "Class-specific deep feature weighting for naïve bayes text classifiers," *IEEE Access*, vol. 8, pp. 20151–20159, 2020, doi: 10.1109/ACCESS.2020.2968984.
[35] A. Luque, M. Mazzoleni, A. Carrasco, and A. Ferramosca, "Visualizing Classification Results: Confusion Star and Confusion Gear," *IEEE Access*, vol. 10, pp. 1659–1677, 2022, doi: 10.1109/ACCESS.2021.3137630.
[36] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.
[37] T. Reineking, "Active classification using belief functions and information gain maximization," *Int. J. Approx. Reason.*, vol. 72, pp. 43–54, May 2016, doi: 10.1016/j.ijar.2015.12.005.
[38] Y. Gao, Y. Feng, and J. Tan, "Exploratory study on cognitive information gain modeling and optimization of personalized recommendations for knowledge reuse," *J. Manuf. Syst.*, vol. 43, pp. 400–408, Apr. 2017, doi: 10.1016/j.jmsy.2017.01.003.
[39] D. Apriliani, T. Abidin, E. Sutanta, A. Hamzah, and O. Somantri, "Sentiment analysis for assessment of hotel services review using feature selection approach based-on decision tree," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 4, pp. 240–245, Jun. 2020, doi: 10.14569/IJACSA.2020.0110432.
[40] F. Gorunescu, *Data Mining: Concepts, Models and Techniques*, vol. 12. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.