SHM PUBLISHER
SURYA HIJAU MANFAAT

Journal of Soft Computing Exploration

Homepage: https://shmpublisher.com/index.php/joscex

JOSCEX

p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

A sentiment analysis of madura island tourism news using C4.5 algorithm

Vina Angelina Savitri[1]*, Moh. Sa'id[2], Husni[3], Arif Muntasa[4]
[1,2,3,4]Department of Informatics Engineering, Universitas Trunojoyo Madura, Indonesia

Article Info

Article history:

Received November 27, 2023
Revised December 22, 2023
Accepted March 4, 2024

Keywords:

Decision tree
Online news article
Sentiment analysis
Madura island tourism

ABSTRACT

Over the past few years, the tourism sector has experienced significant growth in its contribution. The tourism potential on Madura Island is widespread across four regencies, namely Bangkalan, Sampang, Pamekasan, and Sumenep. This potential can be harnessed to support the local government's economy and the communities in the surrounding areas. This research aims to analyze the sentiment of Madura tourism news from online sources using the Decision Tree (C4.5) method. The data used in this study consist of 100 Madura tourism news articles collected from online news portals, which will be classified using the Decision Tree (C4.5) method. The test results show that this method has an average accuracy rate of 76.5% in 10 tests. The average accuracy results demonstrate that the use of the Decision Tree (C4.5) method in this research yields a sufficiently high accuracy level in the sentiment analysis of tourism news.

Corresponding Author:

Vina Angelina Savitri,
Department of Informatics Engineering,
Trunojoyo Madura University,
Email: 190411100170@student.trunojoyo.ac.id

https://doi.org/10.52465/joscex.v5i1.258

1. INTRODUCTION

Madura Island, with an area of approximately 5,025 km2 [1] and a population of around 4,099,070 people [2], boasts vast natural beauty in its four regencies. The island is often referred to as a hidden paradise due to its diverse attractions, including natural, religious, historical, and cultural tourism [3]. Madura Island offers various charming tourist destinations suitable for exploration, making it an extraordinary place to visit.

The tourism sector currently plays a crucial role in supporting Indonesia's economic growth. This is evidenced by data from the Ministry of Tourism, indicating that since 2013, tourism has been the largest foreign exchange earner, ranking fourth after coal, palm oil and oil and gas [4]. Therefore, the use of the tourism sector in Madura Island should be maximized to stimulate economic growth on the island itself. In this regard, local governments are expected to fully support the community's role in the management and development of tourism in Madura Island. One way to attract tourists is through the introduction of tourism through online news media.

Today's easy use of information technology allows technology users to immediately access the desired information and news [5]. Advances in technology can be used as a means to introduce Madura Island tourism. There is a wealth of information and news that has covered Madura Island's tourism, both with positive and negative sentiments. Reviews within the news can provide conclusions about the current conditions at tourist sites [6]. These emotions are expressed in writing and then distributed to the public [7]. Therefore, this sentiment analysis research employs text mining techniques.

Sentiment analysis is a process that applies the text mining method [8]. Text mining is a technique commonly used in text-based information extraction processes. Within text mining, there are several techniques, one of which is sentiment analysis. Sentiment analysis aims to determine the tendency of information, whether it has a positive or negative sentiment, which serves as decision support. The grouping process typically involves the use of classification methods or algorithms, one of which is C4.5.

The IEEE International Conference on Data Mining (ICDM) has identified several best data mining algorithms by categorizing them based on their calculation types. These include algorithms that calculate distances, such as KNN and K-Means, decision tree algorithms like C4.5 and CART, ensemble algorithms like AdaBoost, probability algorithms like Naïve Bayes, and several other algorithms [9]. From the previous explanation, it is evident that C4.5 is considered one of the best data mining algorithms.

Research conducted by Andi Taufik [5] on the classification of hotel reviews applied several methods, including Naïve Bayes, Particle Swarm Optimization (PSO), Support Vector Machine (SVM), and Decision Tree (C4.5). In that study, the Decision Tree (C4.5) method achieved the highest accuracy with 96.94%, while Naïve Bayes obtained an accuracy of 89.98%, SVM model accuracy was 89.86%, and PSO had an accuracy of 95.91%. Another study by Syarifuddin [10] titled “Sentiment Analysis of Public Opinion on the Effects of PSBB on Twitter with Decision Tree-KNN-Naïve Bayes Algorithm” compared various methods. In this research, the Decision Tree method achieved the highest accuracy of 83.3%, while KNN and Naïve Bayes obtained accuracy results of 80.80% and 80.03%, respectively.

Based on the evaluation of previous research, the results indicate that the Decision Tree method (C4.5) has a high level of accuracy compared to several other methods. Therefore, this method is quite efficient in classifying Madura tourism news based on positive and negative sentiments.

2. METHOD

In general, this research is carried out to predict the sentiment tendencies of opinions in tourist news. The study involves several stages of methodology, as outlined in Figure 1. The initial stage of the research is the data collection process, which includes collecting news data from various online sources. The next stage involves labeling the collected dataset, where the news is labeled as positive or negative sentiment. The labeling process is performed manually by the researcher by reading the content of each news article one by one, thus capturing the essence of the news and labeling it with positive or negative sentiment. The purpose of this labeling stage is to train the training data to predict the data in the test set. The next stage is text preprocessing, which has already been implemented using the Python programming language in the process.

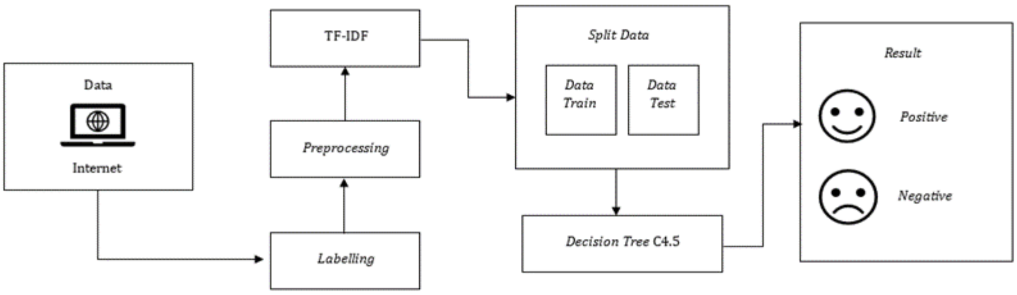

Figure 1. System architecture

Text preprocessing is a data cleaning process that aims to prepare data prior to the classification process. Pre-processing is further divided into several stages, including Case Folding, Filtering, Stopword, Tokenizing, Stemming [11].

1. Case folding is a stage where all capital letters in a news article are converted to lowercase. In news writing, there are usually different letter forms, and the uniformity in letters significantly influences the classification process [12].
2. Filtering, in this stage, irrelevant characters such as emoticons, punctuation, and other unnecessary elements are removed [13].
3. Stopword removal, this stage involves eliminating connector words such as "in," "the,", “with”, and others [14].
4. Tokenizing is the process of dividing a sentence into individual words, commonly known as tokens [15].
5. Stemming is the stage of converting words with affixes into their base form. Stemming is a method based on the morphological rules of the Indonesian language, which classifies affixes as prefixes, infixes, suffixes, and combinations of prefixes and suffixes [16].

The results of preprocessing will then undergo the term frequency-inverse document frequency (TF-IDF) process, where each word in the news will go through a weighting process. The TF-IDF method is used to calculate the relative frequency of each word or token in a document. This method assigns weights to these words based on their importance in the document, considering the number of occurrences of the word in that document and how the word appears in all existing documents [17]. The term frequency (TF) is the calculation of the occurrence of a word in a document divided by the total number of words in the document, as expressed in equation 1.

$$TF = \frac{\alpha}{\beta} \tag{1}$$

Description:
α = The number of occurrences of a word in a document
β = The total number of words in the document
Meanwhile, the inverse document frequency (IDF) is the calculation of the proportion of documents in the corpus that contain the word in the TF calculation. The formula for the calculation of IDF is expressed in equation 2.

$$\mathrm{IDF} = \log\left(\frac{\gamma}{\delta}\right) \tag{2}$$

Description:
γ = The number of occurrences of a word in the document
δ = The total number of words in the document
TF-IDF is the result of multiplying TF by IDF, and the outcome of this multiplication is used as the classification material, as expressed in Equation 3.

$$TF - IDF = TF * IDF \tag{3}$$

The next step is entering the data splitting stage, where the data will be divided into two parts: training data and test data with a ratio of 80% and 20% [18]. The training data are used to train the classification model. After training the classification model with the Decision Tree (C4.5) method using the training data, the classification model will be tested using the test data obtained after the TF-IDF process.

There are many algorithms that can be applied to solve any problem with its own characteristics [19]. One classification technique in text mining is the Decision Tree (C4.5) [20]. Decision Tree is an algorithm that uses a decision tree representation, where attributes are represented as nodes and the branches of the tree represent attribute values, while the class is represented as a label on the node [21]. Decision Tree has several algorithms, namely ID3, C4.5 and CART [22], where the C4.5 flowchart is presented in Figure 2. The C4.5 algorithm is known to exceed the Learning Vector Quantization (LVQ) algorithm with average accuracy and has a fast processing time [23].

The initial step in the C4.5 classification process is to input the data and calculate the entropy value for each attribute using equation 4.

$$Entropy(S) = \sum_{i=1}^{n} -pi * \log_2 pi \tag{4}$$

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy \tag{5}$$

Equation 5 is the formula for calculating the information gain value after entropy calculation. Next, we will make the formation of the tree from the C4.5 classification results. The final step is evaluation, where this process aims to measure the accuracy level of the algorithm used to analyze the sentiment of the news [20]. The accuracy value is the percentage of the data set that can be correctly classified by the system, compared to the overall available news data [24].

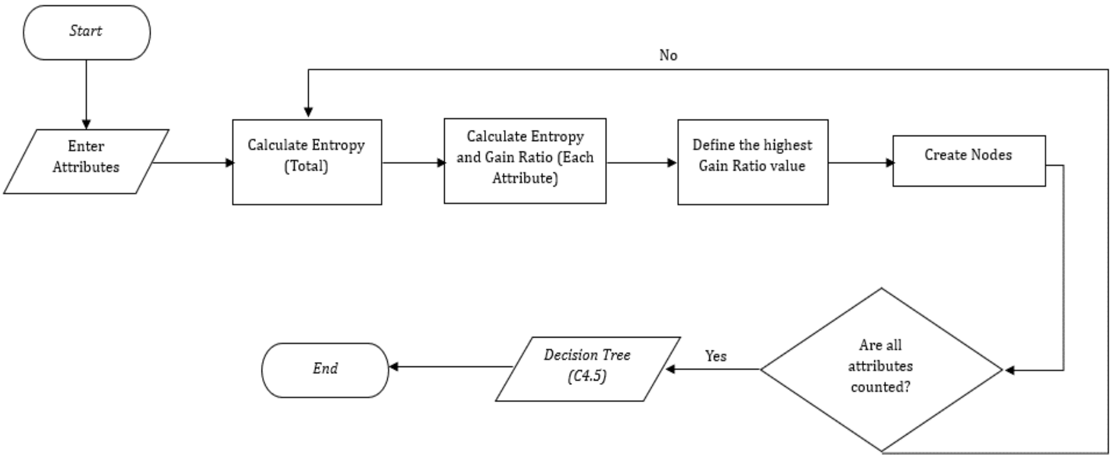
Start
No
Enter Attributes
Calculate Entropy (Total)
Calculate Entropy and Gain Ratio (Each Attribute)
Define the highest Gain Ratio value
Create Nodes
Are all attributes counted?
Yes
Decision Tree (C4.5)
End

Figure 2. Decision Tree (C4.5) flowchart

3. RESULTS AND DISCUSSIONS

The research carried out by the researcher began by collecting news on tourism objects in Madura published from October 2021 to November 2022 from online news sites such as Liputan6.com, Tribunnews.com, Maduraindepth.com, Pamekasanchannel.com, Madurapost.net, Rri.co.id, Beritajatim.com, Portalmadura.com, Opsi.id, Mongabay.co.id, Indozone.id, Detik.com, Koranmadura.com, Okezone.com, Kompas.com, Tvonenews.com, Jatim-times.co.id, Suaraindoneisa-news.com, Maduraya.jurnalisindonesia.id, Idntimes.com, Suarasurabaya.net, Inews.id, Memoonline.co.id, Jatim.suara.com, Radarmadura.jawapost.com, Maduraupdate.com, Jatim.antaranews.com, Maduratoday This research aims to promote tourism destinations and the economy on the island of Madura. The news was manually collected by the researcher by copying the content from the news texts and saving it in CSV format, creating a dataset that will be used in this study.

Data Labeling

Madura Island consists of 4 regions, namely Bangkalan, Sampang, Pamekasan, and Sumenep. From the successfully collected data set, there are 23 news articles discussing tourist attractions in Bangkalan Regency, 13 articles about tourist attractions in Sampang Regency, 27 articles covering tourism-related news in Pamekasan Regency and 37 articles focusing on tourism-related news in Sumenep Regency. This information is illustrated in the diagram shown in Figure 3.

.

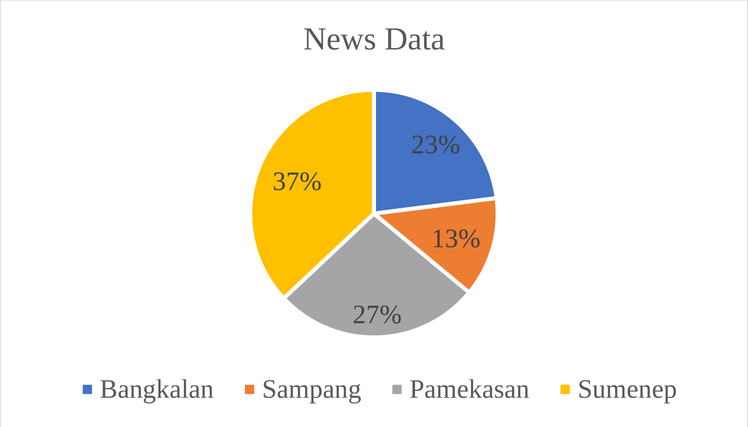

Figure 3. News data

The data set collected in this study consists of 100 news articles discussing tourist attractions in Madura, with the researcher manually labeling the researcher. Of these, 50 news texts are labeled positive sentiment, and the remaining texts are labeled as negative sentiment, as presented in Table 1.

Table 1. Dataset label

Class	Total
Positive	50
Negative	50
Total	100

Text Processing

The collected dataset will be classified using a classification model. However, there are several stages to go through before entering the classification phase. The first stage is pre-processing, which aims to prepare and clean the news text from words and punctuation that are considered less relevant in the classification process. Text pre-processing is further divided into five parts, including case folding, filtering, tokenizing, stopword, and stemming. The results of the preprocessing process are presented in Table 2 with an example sentence, “Hasilnya, perairan Selat Madura terkontaminasi mikroplastik”.

Table 2. Preprocessing result

Steps	After
Case Folding	“hasilnya, perairan selat madura terkontaminasi mikroplastik.”
Filtering	“hasilnya perairan selat madura terkontaminasi mikroplastik”
Tokenizing	[‘hasilnya’, ‘perairan’, ‘selat’, ‘madura’, ‘terkontaminasi’, ‘mikroplastik’]
Stopword	[‘hasilnya’, ‘perairan’, ‘selat’, ‘madura’, ‘terkontaminasi’, ‘mikroplastik’]
Stemming	[‘hasil’, ‘air’, ‘selat’, ‘madura’, ‘kontaminasi’, ‘mikroplastik’]

In the table above, double quotation marks indicate that the sentence within it is a string data type. Meanwhile, square brackets in the tokenizing to stemming stages indicate that the data within them is a list or array data type. The classification process in this study uses the Python programming language to shorten the time and simplify the classification process. In the stemming stage, the Sastrawi library is used to facilitate the Indonesian language. The stemming stage also utilizes Swifter, a Python programming language package designed to accelerate processing time.

TF-IDF and Splitting Data

TF-IDF aims to provide weight and determine the importance of a word in a document. The data from the TF-IDF results will be divided into two parts: training data and test data. The training data serve as the initial data to train the model, while the test data are used to evaluate the model, resulting in accuracy values. The data set is divided into a 80:20 ratio, where 80% of the dataset is used as training data, and the remaining 20% is used as test data. The dynamic data split is facilitated by the train_test_split class available in the scikit-learn library.

Decision Tree C4.5 Method Classification

Finally, the C4.5 decision tree classification process is performed with 10 trials, as illustrated in the graph below. Utilizing the scikit-learn library in the Python programming language greatly aids the classification process because it contains the DecisionTreeClassifier class. This class represents the C4.5 decision tree method, eliminating the need for researchers to code from scratch.

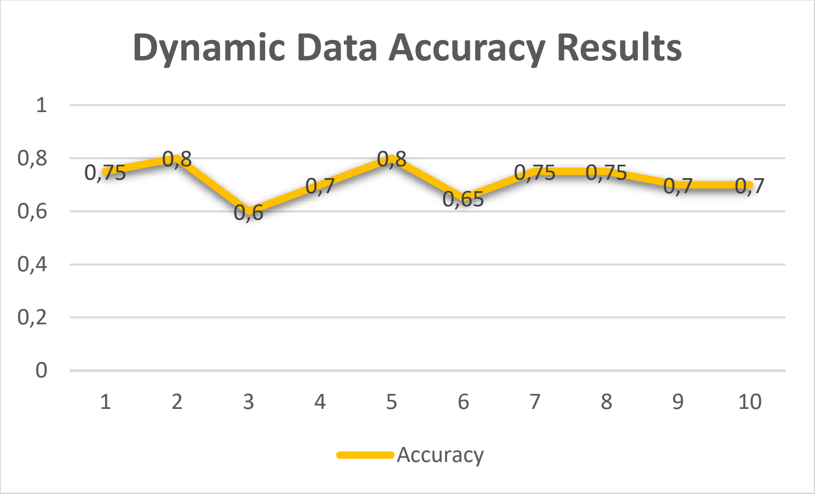

Figure 4. Value dynamic data accuracy

After obtaining the accuracy values of these 10 trials, the average accuracy is found to be 72%. From the accuracy values, it can be interpreted that the Decision Tree C4.5 method provides a sufficiently good accuracy, making it suitable for use in the process of classifying tourism news data. The results of the decision tree from dynamic data splitting are presented in Figure 5. This figure is obtained from experiments with the highest accuracy results, namely 80%, in experiments 2 and 5. The outcome of these experiments reveals the feature with the highest Gini value, which is the feature 'beautiful' with a Gini value of 0.499.

.

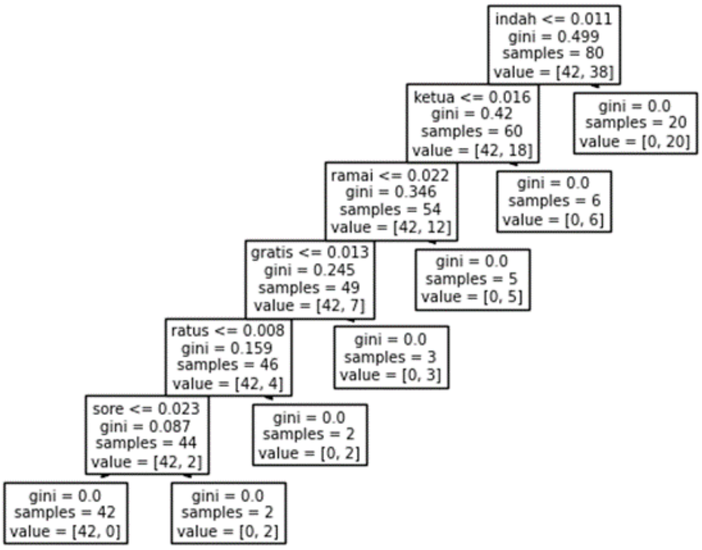

Figure 5. Dynamic data tree results

Furthermore, experiments with static data splitting, where the researcher determined the indices to be used as training and testing data. This experiment was carried out with the aim of obtaining consistent accuracy values. In this experiment, an accuracy result of 80% was achieved with the feature 'beautiful' having the highest Gini value of 0.5. The results from static data splitting decision tree are presented in Figure 6.

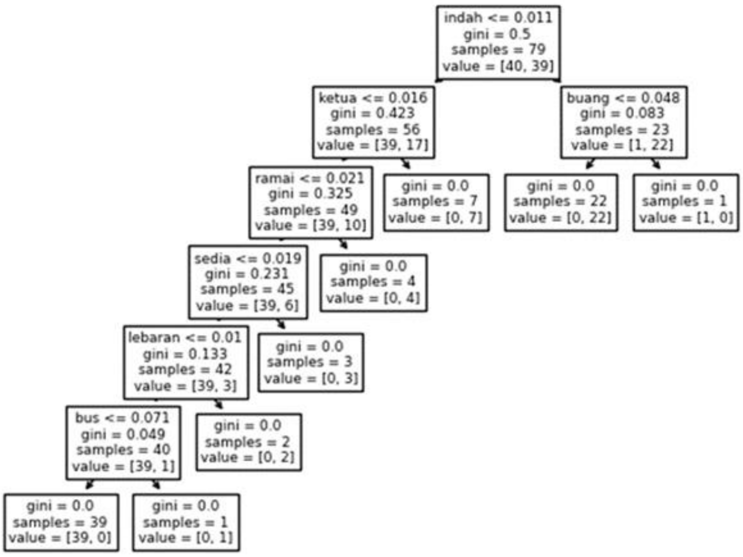

Figure 6. Static data tree results

In this research, a confusion matrix table was obtained to predict the performance of the classification model. Table 3 represents a comparison of the confusion matrix between dynamic data splitting with the highest accuracy and static data splitting.

Table 3. The confusion matrix results from dynamic and static data

	Confusion Matrix	Positive	Negative
Dynamic	Precision	0.79	0.83
	Recall	0.92	0.62
	F1-score	0.85	0.71
Static	Precision	0.88	0.75
	Recall	0.70	0.90
	F1-score	0.78	0.82

Table 3 shows the precision, recall, and f1 score results of the classification model for dynamic and static data splitting. In dynamic data, the precision obtained is 0.83 for negative sentiment and 0.79 for positive sentiment, indicating that the accuracy of the Decision Tree C4.5 method in making predictions is quite high for negative sentiment, likely influenced by the smaller amount of tested data. Furthermore, the recall results yield values of 0.62 for negative sentiment and 0.92 for positive sentiment, indicating that the measurement in identifying all actual cases performed by the model is better for positive sentiment than for negative sentiment. This is likely influenced by the abundance of tested data with positive sentiment. Lastly, the f1 score results, which represent the balance between precision and recall values, are 0.71 for negative sentiment and 0.85 for positive sentiment.

In the static data splitting, the confusion matrix resulted in a precision of 0.75 for data with negative sentiment and 0.88 for data with positive sentiment. Therefore, the model's accuracy in predicting positive sentiment data is higher. The recall values obtained are 0.90 for data with negative sentiment and 0.70 for data with positive sentiment. These results indicate that the errors in recalling actual labels are smaller for negative sentiment data compared to positive sentiment data. Lastly, the f1 score values, representing the balance between precision and recall, are 0.82 for data with negative sentiment and 0.78 for data with positive sentiment. This indicates that the accuracy of the classification model used to group data according to the sentiment present in the content of Madura tourism news texts is quite accurate, as the accuracy values obtained are sufficiently high. This can assist the government in categorizing news data.

The study results showed an accuracy of 80% for static data partitioning and an average accuracy of 72% for dynamic data partitioning in 10 repeated experiments. These findings indicate that the decision tree method is quite efficient in classifying sentiment in tourism news, as it achieved relatively high accuracy. Moreover, the comparison results from the experiments also suggest that static data partitioning performs well compared to dynamic data partitioning.

4. CONCLUSION

This research aimed to address the research gap in understanding sentiment analysis in tourism news. Previously, only a few studies focused on this aspect, and this research seeks to fill that gap. Through the use of the decision tree method, the study achieved an accuracy rate of 80% in static data partitioning and an average accuracy of 72% in dynamic data partitioning over 10 repeated experiments. The research findings indicate that the decision tree method efficiently classifies sentiment in tourism news, achieving a relatively high accuracy level. The comparative results also highlight that static data partitioning performs well, affirming the effectiveness of this approach in the context of this research. The novelty of this research lies in its contribution to our understanding of how the decision tree method can be successfully applied in the sentiment analysis of tourism news. Therefore, this research provides a fresh contribution to the existing literature, paving the way for further studies in this domain.

REFERENCES

[1] Badan Pusat Statistik, “Luas Wilayah,” *Berita Resmi Statistik,* 2023.

[2] Badan Pusat Statistik, “Jumlah Penduduk,” *Berita Resmi Statistik,* 2023.

[3] S. Arifin, “Digitalisasi Pariwisata Madura,” *J. Komun.*, vol. 11, no. 1, p. 53, 2017, doi: 10.21107/ilkom.v11i1.2835.

[4] G. Y. Masyhari Makhasi and S. D. Lupita Sari, “Strategi Branding Pariwisata Indonesia Untuk Pemasaran Mancanegara,” *ETTISAL J. Commun.*, vol. 2, no. 2, p. 31, 2018, doi: 10.21111/ettisal.v2i2.1265.

[5] A. Taufik, “Komparasi Algoritma Text Mining Untuk Klasifikasi Review Hotel,” *J. Tek. Komput.*, vol. IV, no. 2, pp. 112–118, 2018, doi: 10.31294/jtk.v4i2.3461.

[6] L. R. Putri, “Pengaruh Pariwisata Terhadap Peningkatan Kota Surakarta,” *Cakra Wisata*, vol. 21, no. 1, pp. 43–49, 2020.

[7] J. Jumanto, M. A. Muslim, Y. Dasril, and T. Mustaqim, “Accuracy of Malaysia Public Response to Economic Factors During the Covid-19 Pandemic Using Vader and Random Forest,” *J. Inf. Syst. Explor. Res.*, vol. 1, no. 1, pp. 49–70, 2022.

[8] A. Falasari and M. A. Muslim, “Optimize Naïve Bayes Classifier Using Chi Square and Term Frequency Inverse Document Frequency For Amazon Review Sentiment Analysis,” *J. Soft Comput. Explor.*, vol. 3, no. 1, pp. 31–36, Mar. 2022, doi: 10.52465/joscex.v3i1.68.

[9] A. H. Nasrullah, “Implementasi Algoritma Decision Tree Untuk Klasifikasi Produk Laris,” *J. Ilm. Ilmu Komput.*, vol. 7, no. 2, pp. 45–51, 2021, doi: 10.35329/jiik.v7i2.203.

[10] M. Syarifuddin, “ANALISIS SENTIMEN OPINI PUBLIK TERHADAP EFEK PSBB PADA TWITTER DENGAN ALGORITMA DECISION TREE-KNN-NAÏVE BAYES,” vol. 15, no. 1, pp. 87–94, 2020, doi: 10.33480/inti.v15i1.1433.

[11] A. I. Kadhim, “An Evaluation of Preprocessing Techniques for Text Classification,” *Int. J. Comput. Sci. Inf. Secur.*, vol. 16, no. 6, pp. 22–32, 2018.

[12] O. Somantri and D. Dairoh, “Analisis Sentimen Penilaian Tempat Tujuan Wisata Kota Tegal Berbasis Text Mining,” *J. Edukasi*

dan Penelit. Inform., vol. 5, no. 2, p. 191, 2019, doi: 10.26418/jp.v5i2.32661.

[13] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, “Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm,” *Procedia Comput. Sci.*, vol. 161, pp. 765–772, 2019, doi: 10.1016/j.procs.2019.11.181.

[14] S. Kannnan and V. Gurusamy, “Preprocessing Techniques for Text Mining,” *Int. J. Comput. Sci. Commun. Networks*, no. February 2015, pp. 7–16, 2014.

[15] A. Murugan, H. Chelsey, and N. Thomas, *Practical Text Analytics*, vol. 37. 2019.

[16] N. Yusliani, R. Primartha, and M. Diana, “Multiprocessing Stemming: A Case Study of Indonesian Stemming,” *Int. J. Comput. Appl.*, vol. 182, no. 40, pp. 15–19, 2019, doi: 10.5120/ijca2019918476.

[17] S.-W. Kim and J.-M. Gil, “Research paper classification systems based on TF-IDF and LDA schemes,” *Human-centric Comput. Inf. Sci.*, vol. 9, no. 1, p. 30, Dec. 2019, doi: 10.1186/s13673-019-0192-7.

[18] C. Tang, D. Wang, A. H. Tan, and C. Miao, “EEG-Based Emotion Recognition via Fast and Robust Feature Smoothing,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10654 LNAI, no. November, pp. 83–92, 2017, doi: 10.1007/978-3-319-70772-3_8.

[19] A. F. Mulyana, W. Puspita, and J. Jumanto, “Increased accuracy in predicting student academic performance using random forest classifier,” *J. Student Res. Explor.*, vol. 1, no. 2, pp. 94–103, Jul. 2023, doi: 10.52465/josre.v1i2.169.

[20] H. H. Patel and P. Prajapati, “Study and Analysis of Decision Tree Based Classification Algorithms,” *Int. J. Comput. Sci. Eng.*, vol. 6, no. 10, pp. 74–78, 2018, doi: 10.26438/ijcse/v6i10.7478.

[21] S. Wahyuni, “Implementation of Data Mining to Analyze Drug Cases Using C4.5 Decision Tree,” *J. Phys. Conf. Ser.*, vol. 970, no. 1, 2018, doi: 10.1088/1742-6596/970/1/012030.

[22] J. Mantik, N. Abdillah, and M. Ihksan, “Application of the C4 . 5 Algorithm for Classification of Medical Record Data At M . Djamil Hospital Based on the International Disease Code,” vol. 6, no. 36, pp. 576–581, 2022.

[23] A. Nazir, A. Akhyar, Y. Yusra, and E. Budianita, “Toddler Nutritional Status Classification Using C4.5 and Particle Swarm Optimization,” *Sci. J. Informatics*, vol. 9, no. 1, pp. 32–41, May 2022, doi: 10.15294/sji.v9i1.33158.

[24] D. Saputra, W. Irmayani, D. Purwaningtias, and J. Sidauruk, “A Comparative Analysis of C4.5 Classification Algorithm, Naïve Bayes and Support Vector Machine Based on Particle Swarm Optimization (PSO) for Heart Disease Prediction,” *Int. J. Adv. Data Inf. Syst.*, vol. 2, no. 2, pp. 84–95, 2021, doi: 10.25008/ijadis.v2i2.1221.