Jurnal Bumigora Information Technology (BITe)
Vol. 5, No. 2, Desember 2023, Hal. 151∼158
E-SSN: 2685-4066, DOI: 10.30812/bite/v5i1.3166

Performance Improvement of The Random Forest Method Based on Smote-Tomek Link on Lombok Tourism Analysis Sentiment

Khairan Marzuki[1*], Lalu Ganda Rady Putra[2], Hairani Hairani[3], Lalu Zazuli Azhar Mardedi[4]
Juvinal Ximenes Guterres[5]
[1,2,3,4]Universitas Bumigora, Mataram, Indonesia
[5]Universidade Oriental Timur Lorosae, Unital Becora Dili, Timor Leste
Khairan.marzuki@universitasbumigora.ac.id[1*], lalugandaradyputra@universitasbumigora.ac.id[2],
hairani@universitasbumigora.ac.id[3], pujutsega@gmail.com[4], guterresmenex@gmail.com[5]

Article Info:
Diterima: 30 September 2023, Direvisi: 02 November 2023, Disetujui: 31 Desember 2023

Abstract-
Background: Tourists visiting the Lombok island can access various sources of tourist information and can share their views and tourist experiences through social media such as positive and negative experiences.
Objective: This research aims to analyze the sentiment of Lombok tourism reviews using the Smote-Tomek Link and Random Forest algorithms.
Methods: The research was carried out in several stages, namely collecting the Lombok tourism dataset, text preprocessing, text weighting using the Term Frequency-Inverse Document Frequency (TF-IDF) method, data sampling using SMOTE-Tomek Link, text classification using Random Forest, and the final stage was performance testing based on accuracy.
Result: The research results obtained using the Smote-Tomek Link and Random Forest methods in sentiment analysis of tourist reviews about Lombok were 94%.
Conclusion: The use of the Smote-Tomek Link and Random Forest methods in Lombok tourism sentiment analysis produces very good accuracy.

Keywords: Lombok Tourism, Smote-Tomek Link, Random Forest, Improvement Performance

Correspondence Author:
Khairan Marzuki,
Department of Computer Science, Universitas Bumigora, Mataram, Indonesia
Email: khairan.marzuki@universitasbumigora.ac.id

How to Cite: K. Marzuki, L. G. R. Putra, Hairani, L. Z. A. Mardedi, and J. X. Guterres, "Performance Improvement of The Random Forest Method Based on Smote-Tomek Link on Lombok Tourism Analysis Sentiment", *Jurnal Bumigora Information Technology (BITe)*, vol. 5, no. 2, pp. 151∼158, 2023.

1. INTRODUCTION

Tourism is one of the country's assets that plays an important role in improving the economy of a region. Tourism is also able to be a factor for the country's economic growth and development. In Indonesia, tourism has become a mainstay and potential sector to be developed. No wonder President Joko Widodo's administration has made the tourism sector one of the priority programmes to improve the Indonesian economy. One of the areas in Indonesia that is famous for its natural beauty such as beaches is Lombok Island. Not only beaches, Lombok also has hills, mountains, waterfalls, and traditional villages that are no less interesting to visit. One of the most visited tourist attractions by tourists is the area in Mandalika. Mandalika area tourism has become a favourite place to be visited by tourists, because it provides many tourist options such as Kuta Beach, Tanjung Aan Beach, Mawun Beach, Seger Beach, Bukit Merese, Sade Village, and Mandalika Circuit.

Travellers now visiting the island can access a variety of tourism information sources and can share their views and experiences. Tourism content shared through social media has become a highly influential source of information that impacts tourism in terms of both reputation and performance. However, the volume of data on the Internet has reached a level that makes manual processing almost impossible, thus demanding new analytics approaches. Sentiment analysis is rapidly emerging as an automated process for examining semantic relationships and meanings in reviews. Technological advances have fundamentally changed how information is produced and consumed by all actors involved in tourism. Therefore, it is necessary to analyse the sentiment of tourists towards their experience of Lombok tourism, especially the mandalika tourist area in order to find out positive and negative sentiments.

Some previous related research that has done sentiment analysis of tourism is like research [1] using data mining and LDA methods for modelling the sentiment of Malaysian people towards tourism in the country. The accuracy of the method used is 86%. Research [2] uses SVM and Naive Bayes methods for sentiment analysis of Saudi Arabian tourism based on community tweets with SVM method accuracy results of 85% and Naive Bayes 82%. Research [3] using the CNN-LSTM method for the classification of educational tourism reviews with an accuracy of 91%. Research [4] used the KNN method for sentiment analysis of Madura tourism with an accuracy of 94%. Research [5] using SVM, KNN, Naive Bayes, and Decision Tree methods for sentiment analysis of Bali tourism. Based on the results of his research, the best method accuracy result is the SVM method of 81%.

Research [6] using the LSTM method for sentiment analysis of tourist attractions on Trip ADVISOR with an accuracy result of 96%. Research [7] using SVM, Random Forest, and RNN LSTM methods. Based on the results of his research, the RNN LSTM method gets the best accuracy compared to the two methods used by 81%. Research [8] using SVM, Random Forest, and CART methods for sentiment analysis of Thai tourism. Based on the results of his research, the Random Forest method gets the best accuracy compared to the two methods used by 95.4%.Research [9] uses the AHP - SAW method for selecting the best tour. Research [**?**] uses the Naive Bayes method for sentiment analysis of Lombok tourism with 92% accuracy. Research [11] uses the KNN method for sentiment analysis of villa reviews in Ubud with an accuracy of 91%. Research [12] uses the Naive Bayes method for tourism sentiment analysis in the Covid period with an accuracy of 62%. Research [13] uses the LSTM method for sentiment analysis of bali tourism with an accuracy of 96%.

Some previous studies have weaknesses that can be improved in this study, namely that previous studies have not solved the problem of imbalance in the dataset used. Solving the problem of unbalanced datasets on tourism review data needs to be done in order to improve the performance of the method used [14, 15]. Unbalanced data is the amount of data in one class more than in other classes. The problem of data imbalance causes the classification method to classify the majority class more dominantly than the minority class, or in other words, the classification method ignores the minority class. The problem of unbalanced data can be handled with a data sampling approach.

Based on the description above, the solution offered in this research is to use the Smote-Tomek Link over-sampling approach to balance the data so as to improve the performance of the classification method used. This research uses the Random Forest classification method for sentiment analysis of tourism in Lombok, especially

in the Mandalika tourism area. The Random Forest method was chosen because it has the advantages of high accuracy, handling noise data, fast performance in training data, overfitting control, and easy implementation [16]. **The purpose of this research** is to conduct sentiment analysis of Lombok tourism reviews using the Smote-Tomek Link and Random Forest algorithms.

2. METHODS

The stages of research used in this study are shown in Figure 1.

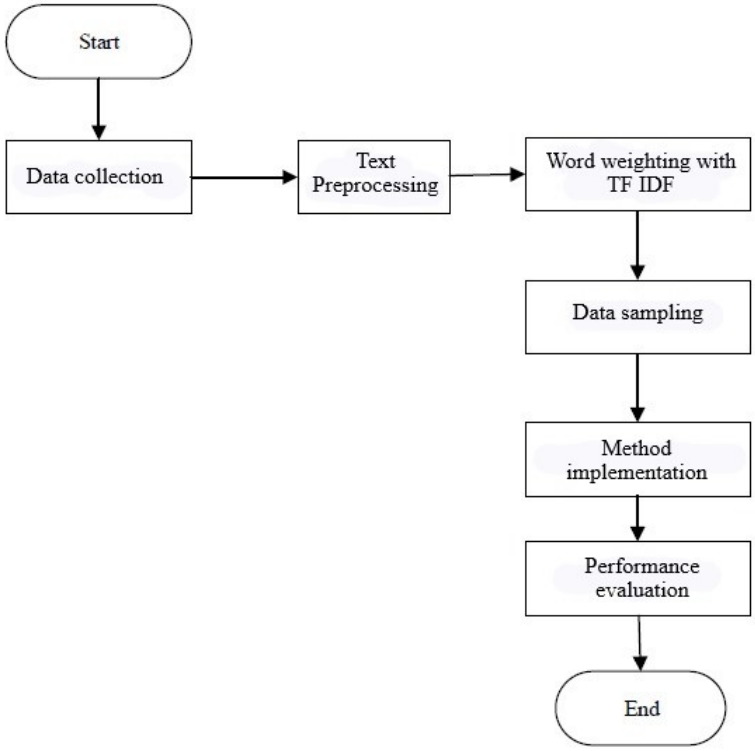

Figure 1. The research flow

Based on Figure 1, this research consists of several stages, namely collecting datasets by scraping twiter user reviews about Lombok tourism, especially the mandalika area. Next is the labelling of review data conducted by Indonesian language experts which is divided into positive, neutral, and negative reviews. The next stage is the preprocessing stage. In the text preprocessing section, text cleaning is used such as case folding, tokenisation, filtering, and stemming. The results of text preprocessing are then carried out text weighting using the TF-IDF method. The next stage is data sampling using the SMOTE-Tomek Link method to balance the data. After the data is balanced, text classification is carried out using the Random Forest method based on the division of training data by 80% and testing by 20%. The last stage is performance testing based on accuracy using the confusion matrix table.

2.1. Dataset Collecting

The data used are twiter user reviews about Lombok tourism, especially the mandalika area obtained by textitscraping data. The results of textitscraping user review data do not yet contain review categories such as positive, neutral, and negative.

2.2. Preprocessing Texts

The text preprocessing stage is a process of changing the form of data to be more structured according to its needs in the data mining process. The stages commonly used in the text preprocessing stage are shown in Figure 2.

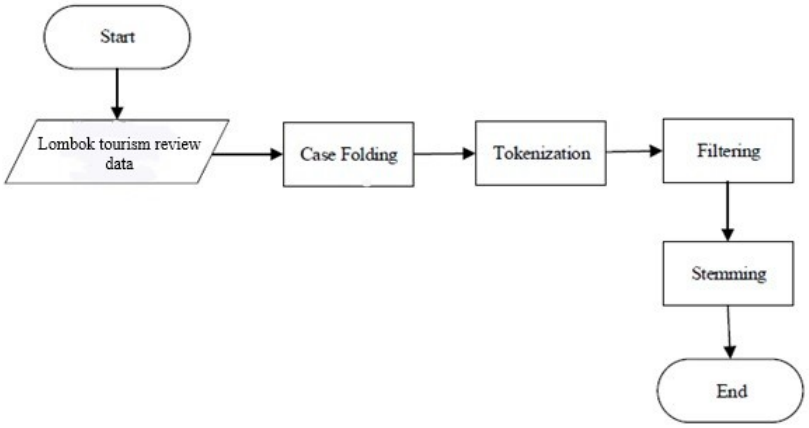

Figure 2. Text Preprocessing Stages

In Figure 2 the preprocessing stage begins with case folding. Case folding is the process of changing capital letters into lowercase letters. Tokenisation is the process of separating sentences into words. Filtering is used to discard less important words or keep important words. Common words that usually appear in Indonesian are "yang," "dan," "di," "dari.". Stemming is the process of forming base words.

2.3. Data Weighting

The preprocessing data that is still in the form of words will be converted into numbers with a word weighting process that aims to calculate the weight on each word that will be used as a feature. The result of word weighting with TF-IDF is the multiplication of TF and IDF values which will produce the weight of each word using Equation (1). Tf is term frequency, W is TF-IDF weight, and idf is Inverse document Frequency.

$$W = tf \times idf \tag{1}$$

2.4. Data Sampling

Unbalanced data needs to be addressed in order to improve the performance of the classification method used. This research uses the SMOTE-Tomek Link method to balance the data. The smote method works by adding instances to the minority class based on the nearest neighbour. The stages of the SMOTE-Tomek Link method in balancing the data are shown in Figure 3. Tomek Link is an undersampling method that cleans noise data from majority classes that have similar and overlapping characteristics. Tomeklink works by removing majority class instances that are closer to the minority class by applying the nearest neighbour rule to select instances. The combination of Tomeklink and Smote oversampling can improve accuracy better than individual performance [17].

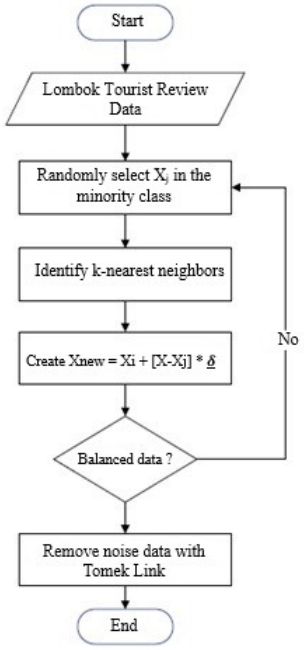

Figure 3. Smote-Tomek Link Flow

2.5. Model Implementation

At this stage, the implementation of the algorithm used in classification will be carried out, namely the Random Forest method. Random Forest is a decision tree-based ensemble learning method [18] which has advantages such as high accuracy, ability to handle noise data, fast performance in training data, overfitting control, and easy to implement [16]. The Random Forest algorithm works by creating a set of decision trees from a randomly selected subset, obtaining predictions from each decision tree, voting for each predicted result, and selecting the best predicted result based on the most votes set as the final prediction.

2.6. Work Evaluation

The performance of the random forest algorithm is measured based on accuracy using a confusion matrix table. Confusion matrix is a table that describes the performance of a classification method on a dataset whose true values are known. Confusion matrix can visualise the number of correctly and incorrectly classified data as shown in Table 1. The accuracy calculation uses Equation (2).

Table 1. Confusion Matrixs

Actual	Prediction	
	Negatif	Positif
Negatif	TN	FP
Positif	TF	TP

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

3. RESULTS AND DISCUSSION

This section is used to describe the research results that have been obtained at each stage. The first stage is the collection of datasets related to the topic of Lombok tourism, especially Mandalika tourism through scraping data on Twitter with the help of python tools. The amount of data related to Lombok tourism is 230 data that does not yet have a label. Datasets of 230 data are labelled by Indonesian language experts into 3 categories of positive, negative, and neutral. The Lombok tourism topic data that already has a label is shown in Table 2.

Table 2. Lombok Tourism Topic Dataset

No	Tweet	Label
1	1. Pantai Seger Kuta Tak hanya Bali, Lombok juga punya Pantai Kuta, lho. Pantai ini merupakan bagian dari kawasan wisata Mandalika yang berada di Lombok. https://t.co/FmLeeaSdyv	Positif
2	Amazing Indonesia spesial Mandalika ditutup dengan Pantai Kuta Mandalika, Tanjung Aan hingga Bukit Merese yang menjadi wisata tak terlupakan bagi para penikmat langit dan lautan. #TVRI #TVRINasional #MediaPemersatuBangsa #KutaMandalika #TanjungAan #Mandalika #Indonesia https://t.co/NYW83ZFEuq	Positif
.....	.....	.....
229	Wisatawan Mengeluh Tarif Masuk Pantai Seger, Saber Pungli Diterjunkan: https://t.co/6LhQoAjAmc Dunia pariwisata di Lombok Tengah kembali tercoreng karena maraknya pungutan liar atau pungli saat masuk di Pantai Seger, The post Wisatawan Mengeluh Tarif https://t.co/Ph5Iq9t5w8 https://t.co/s3gd251032	Negatif
230	Sorenya kita ke Bukit Merese, emang ga lengkap kalo ke Lombok ga ke Bukit Merese. Viewnya 2 pantai, anginnya kenceng banget! https://t.co/Mpj3r9575L	Negatif

The data in Table 2 is still raw or unqualified so it is necessary to process the data in order to improve the performance of the classification method. This research uses several preprocessing techniques such as case folding, tokenisation, filtering, and stemming. The case folding process changes capital letters into lowercase letters. Tokenisation is the process of separating sentences into words. Filtering is used to discard less important words or keep important words. Common words that usually appear in Indonesian are "yang," "dan," "di," "dari." Stemming is the process of forming base words. The text that has been processed can be shown in Table 3.

Table 3. Dataset After Pre-processing

No	Original Data	Results of Pre-processing Data	Label
1	1. Pantai Seger Kuta Tak hanya Bali, Lombok juga punya Pantai Kuta, lho. Pantai ini merupakan bagian dari kawasan wisata Mandalika yang berada di Lombok. https://t.co/FmLeeaSdyv	['pantai', 'seger', 'kuta', 'bali', 'lombok', 'pantai', 'kuta', 'lho', 'pantai', 'kawasan', 'wisata', 'mandalika', 'lombok']	Positif
2	Amazing Indonesia spesial Mandalika ditutup dengan Pantai Kuta Mandalika, Tanjung Aan hingga Bukit Merese yang menjadi wisata tak terlupakan bagi para penikmat langit dan lautan. #TVRI #TVRINasional #MediaPemersatuBangsa #KutaMandalika #TanjungAan #Mandalika #Indonesia https://t.co/NYW83ZFEuq	['amazing', 'indonesia', 'spesial', 'mandalika', 'tutup', 'pantai', 'kuta', 'mandalika', 'tanjung', 'aan', 'bukit', 'merese', 'wisata', 'lupa', 'nikmat', 'langit', 'laut']	Positif
.....	.....	.....	.....
229	Wisatawan Mengeluh Tarif Masuk Pantai Seger, Saber Pungli Diterjunkan: https://t.co/6LhQoAjAmc Dunia pariwisata di Lombok Tengah kembali tercoreng karena maraknya pungutan liar atau pungli saat masuk di Pantai Seger, The post Wisatawan Mengeluh Tarif https://t.co/Ph5Iq9t5w8 https://t.co/s3gd251032	['wisatawan', 'keluh', 'tarif', 'masuk', 'pantai', 'seger', 'saber', 'pungli', 'terjun', 'dunia', 'pariwisata', 'lombok', 'coreng', 'marak', 'pungut', 'liar', 'pungli', 'masuk', 'pantai', 'seger', 'the', 'post', 'wisatawan', 'keluh', 'tarif]	Negatif
230	Sorenya kita ke Bukit Merese, emang ga lengkap kalo ke Lombok ga ke Bukit Merese. Viewnya 2 pantai, anginnya kenceng banget! https://t.co/Mpj3r9575L	['sore', 'bukit', 'merese', 'emang', 'lengkap', 'kalo', 'lombok', 'bukit', 'merese', 'viewnya', 'pantai', 'angin', 'kenceng', 'banget']	Negatif

Clean text data resulting from preprocessing is weighted using the TF-IDF method. Weighting is done to convert words or terms to numeric so that classification can be done using the random forest method. Each term or word has a weight that represents the level of importance. Data that has been given weight, data balancing is done using Smote-Tomek Link. The results of data balancing using Smote-Tomek Link are shown in Table 4.

Data that has been balanced using the Smote-Tomek Link, then the classification process is carried out using the Random Forest method. Training data is used by 80% and testing data by 20%. **The findings of this study are** that the Smote-Tomek Link method with Random Forest provides very high accuracy in analysing tourist opinions on Lombok tourism, where the accuracy obtained is 94% based on the table of confusion matrix

Table 4. Data Distribution Before and After Smote-Tomek Link

Class	Original Data	Results of Smote-Tomek Link
Positif	109	102
Netral	105	101
Negatif	16	109

shown in Table 5. **The results of this study are** in line with the research of [14, 19] which obtained very good accuracy by applying the Smote-Tomek Link method to balance the data before classification by the classification method used.

Table 5. Confusion Matrix Result of Random Forest Method

Actual	Prediction		
	Negatif	Netral	Positif
Negatif	25	0	0
Netral	0	16	3
Positif	0	1	18

In Table 3, the Smote-Tomek Link with Random Forest method successfully classified the negative class correctly for 25 instances, the neutral class was classified for 16 instances out of a total of 19 instances, and the positive class was correctly classified for 18 instances out of a total of 19 instances.

4. CONCLUSSION

It has been successfully implemented the Smote-Tomek Link and Random Forest methods on tourism sentiment analysis in Mandalika Lombok with 3 categories of reviews such as Positive, Neutral, and Negative. The use of the Smote-Tomek Link and Random Forest methods in the sentiment analysis of Lombok tourism resulted in excellent accuracy of 94%. Future research can use the LDA method to describe the topics discussed by tourists about Lombok tourism, especially Mandalika tourism.

ACKNOWLEDGEMENT

Our sincere thanks to DRTPM Kemendikbudristek for providing funding for this research so that this article can be published.

REFERENCES

[1] N. A. Deraman, A. G. Buja, K. A. F. A. Samah, M. N. H. H. Jono, M. A. M. Isa, and S. Saad, "A social media mining using topic modeling and sentiment analysis on tourism in Malaysia during COVID19," *IOP Conference Series: Earth and Environmental Science*, vol. 704, no. 1, pp. 1–9, 2021.

[2] S. M. Alrashidi and A. M. Awadelkarim, "Machine Learning-Based Sentiment Analysis for Tweets Saudi Tourism," *Journla of Theoretical and Applied Information Technology*, vol. 100, no. 16, pp. 5096 –5109, 2022.

[3] Y. Wang, C. Chu, and T. Lan, "Sentiment Classification of Educational Tourism Reviews Based on Parallel CNN and LSTM with Attention Mechanism," *Mobile Information System*, pp. 1–13, 2022.

[4] F. H. Rachman, Imamah, and B. S. Rintyarna, "Sentiment Analysis of Madura Tourism in New Normal Era using Text Blob and KNN with Hyperparameter Tuning," in *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, jan 2022, pp. 23–27.

[5] C. Steven and W. Wella, "The Right Sentiment Analysis Method of Indonesian Tourism in Social Media Twitter," *IJNMT (International Journal of New Media Technology)*, vol. 7, no. 2, pp. 102–110, 2020.

[6] N. Hanafiah, Y. Setiawan, A. Buntaran, and M. Reynaldi, "Sentiment Analysis of Tourism Objects on Trip Advisor Using LSTM Method," *Journal of Computer Science and Technology Studies*, vol. 4, no. 2, pp. 1–6, 2022.

[7] R. K. Mishra, S. Urolagin, J. A. A. Jothi, A. S. Neogi, and N. Nawaz, "Deep Learning-based Sentiment Analysis and Topic Modeling on Tourism During Covid-19 Pandemic," *Frontiers in Computer Science*, vol. 3, no. November, pp. 1–14, 2021.

[8] N. Leelawat, S. Jariyapongpaiboon, A. Promjun, S. Boonyarak, K. Saengtabtim, A. Laosunthara, A. K. Yudha, and J. Tang, "Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning," *Heliyon*, vol. 8, no. 10, p. e10894, 2022. [Online]. Available: https://doi.org/10.1016/j.heliyon.2022.e10894

[9] A. I. J. Nisa, R. Prawiro, and N. Trisna, "Analisis Hybrid DSS untuk Menentukan Lokasi Wisata Terbaik," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 2, pp. 238–246, 2021.

[10] N. L. P. M. Putu, A. Z. Amrullah, and Ismarmiaty, "Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma Naive Bayes dan Latent Dirichlet Allocation," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 123–131, 2021.

[11] N. L. W. S. R. Ginantra, C. P. Yanti, G. D. Prasetya, I. B. G. Sarasvananda, and I. K. A. G. Wiguna, "Analisis Sentimen Ulasan Villa di Ubud Menggunakan Metode Naive Bayes, Decision Tree, dan K-NN," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 11, no. 3, pp. 205–215, 2022.

[12] D. Arsa, I. Weni, and A. Fahreza, "Analisis Sentimen Terhadap Pariwisata di MasaCovid-19 Menggunakan Naïve Bayes," *Jurnal Telematika*, vol. 17, no. 1, pp. 49–54, 2022.

[13] D. I. Af'idah, D. Dairoh, S. F. Handayani, R. W. Pratiwi, and S. I. Sari, "Sentimen Ulasan Destinasi Wisata Pulau Bali Menggunakan Bidirectional Long Short Term Memory," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 607–618, 2022.

[14] H. Hairani, A. Anggrawan, and D. Priyanto, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link," *International Journal on Informatics Visualization*, vol. 7, no. 1, pp. 258–264, 2023.

[15] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89–93, jan 2020.

[16] K. Guo, X. Wan, L. Liu, Z. Gao, and M. Yang, "Fault diagnosis of intelligent production line based on digital twin and improved random forest," *Applied Sciences (Switzerland)*, vol. 11, no. 16, jan 2021.

[17] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset," *Sensors*, vol. 22, no. 9, jan 2022.

[18] Y. Sun, H. Zhang, T. Zhao, Z. Zou, B. Shen, and L. Yang, "A New Convolutional Neural Network with Random Forest Method for Hydrogen Sensor Fault Diagnosis," *IEEE Access*, vol. 8, pp. 85 421–85 430, 2020.

[19] I. N. Switrayana, D. Ashadi, H. Hairani, and A. Aminuddin, "Sentiment Analysis and Topic Modeling of Kitabisa Applications using Support Vector Machine (SVM) and Smote-Tomek Links Methods," *International Journal of Engineering and Computer Science Applications (IJECSA)*, vol. 2, no. 2, pp. 81–91, jan 2023.