JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7472

DETECTION OF MICRO-VIRAL CONTENT ON TIKTOK THROUGH SOCIAL LISTENING AND MACHINE LEARNING

Ratih Anggraeni[1]; Purwadi[2]*; Pungkas Subarkah[1]

Department of Informatics[1]
Master of Computer Science[2]
Universitas Amikom Purwokerto, Purwokerto, Indonesia[1,2]
https://amikompurwokerto.ac.id[1,2]
22sa11a152@amikompurwokerto.ac.id, purwadi@amikompurwokerto.ac.id*,
subarkah@amikompurwokerto.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *The phenomenon of micro-virality on TikTok illustrates how content can rapidly spread on a small scale before reaching broader virality. Understanding its driving factors is essential for supporting digital marketing strategies, managing content creators, and analyzing social media trends. This study aims to detect and predict the potential of micro-virality in TikTok videos by integrating a social listening approach with machine learning techniques. The dataset consists of approximately 4,000 TikTok posts enriched with 20 features across five categories, including user metadata (author popularity, follower ratio), temporal features (posting time and day), network features (hashtags and mentions), content features (text length and keywords), and contextual elements (trending music and video duration). To ensure objective labeling, a quantile-based threshold was applied, categorizing videos in the top 25% of view counts (≥ 26,300,000 views) as viral, resulting in a class distribution of 24.74% viral and 75.26% non-viral. To address this imbalance, the SMOTENC technique was used to oversample the minority class and enhance data representativeness. Three machine learning algorithms Random Forest, Extreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN) were implemented. Experimental results show that Random Forest improved from 88% to 92%, XGBoost maintained strong performance at 95%, and ANN increased significantly from 92% to 93% after SMOTENC application. These findings indicate that SMOTENC effectively improves model generalization and reduces bias toward majority classes, supporting more reliable early-stage virality prediction. Overall, the study enriches social media analytics research and provides practical insights for optimizing TikTok content strategies and early trend detection.*

Keywords: *Machine Learning, Micro-Virality, Smotenc, Social Listening, TikTok Content Analytics.*

Intisari— *Fenomena mikro-viralitas di TikTok menggambarkan bagaimana suatu konten dapat menyebar dengan cepat pada skala kecil sebelum mencapai tingkat viral yang lebih luas. Memahami faktor-faktor yang mendorong penyebaran awal ini penting untuk mendukung strategi pemasaran digital, manajemen kreator, dan analisis tren media sosial. Penelitian ini bertujuan untuk mendeteksi dan memprediksi potensi mikro-viralitas pada video TikTok dengan mengintegrasikan pendekatan social listening dan teknik machine learning. Dataset terdiri dari sekitar 4.000 unggahan TikTok yang diperkaya dengan 20 fitur dari lima kategori, yaitu metadata pengguna (popularitas kreator, rasio pengikut), fitur temporal (waktu dan hari unggah), fitur jejaring (hashtags dan mentions), fitur konten (panjang teks dan kata kunci), serta elemen kontekstual (musik tren dan durasi video). Untuk memastikan pelabelan yang objektif, threshold viral ditentukan menggunakan metode kuantil, di mana video yang berada pada 25% teratas jumlah penayangan (≥ 26.300.000 penayangan) diklasifikasikan sebagai viral, menghasilkan distribusi kelas 24,74% viral dan 75,26% non-viral. Teknik SMOTENC kemudian diterapkan untuk mengatasi ketidakseimbangan kelas dengan melakukan oversampling pada kelas minoritas. Tiga algoritma machine learning digunakan, yaitu Random*

VOL. 11. NO. 3 FEBRUARY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i3.7472

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

Forest, Extreme Gradient Boosting (XGBoost), dan Artificial Neural Network (ANN). Hasil eksperimen menunjukkan bahwa Random Forest meningkat dari 88% menjadi 92%, XGBoost mempertahankan performa kuat pada 95%, dan ANN meningkat signifikan dari 92% menjadi 93% setelah penerapan SMOTENC. Temuan ini menunjukkan bahwa SMOTENC secara efektif meningkatkan generalisasi model dan mengurangi bias terhadap kelas mayoritas, sehingga menghasilkan prediksi mikro-viralitas yang lebih andal. Secara keseluruhan, penelitian ini memperkaya kajian analitik media sosial serta memberikan wawasan praktis untuk optimalisasi strategi konten TikTok dan deteksi tren awal.

***Kata Kunci**: Pembelajaran Mesin, Mikro-Viralitas, Smotenc, Social Listening, Analitik Konten Tiktok*

INTRODUCTION

TikTok has grown into one of the most influential social media platforms with over one billion monthly active users worldwide. The platform is not only used as a means of entertainment, but also as an important medium for digital marketing, information dissemination, and shaping public opinion [1]. One interesting phenomenon on TikTok is micro-virality, which refers to small-scale virality occurring within a specific community or segment before expanding into broader popularity. This phenomenon has a significant impact on marketing strategies, content creator management, and social trend analysis [2]. In this study, micro-viral content is defined as posts which reaches over 100,000 views in the first 24 hours, represents a measurable threshold that differentiates early-stage viral content from non-viral posts..

Several previous studies have explored factors influencing the virality of social media content. For instance, Krowinska A. (2025) found that early interactions such as likes, comments, and shares are key indicators of potential virality [3]. Keir A. (2021) emphasized the role of hashtags as connectors between communities that accelerate content dissemination [4]. Meanwhile, Rajapaksha R. (2023) highlighted how video duration and local cultural context influence audience perceptions of content popularity [5]. In addition, limited research has examined micro-virality as a distinct early-stage phenomenon rather than a subset of overall virality. Existing studies also tend to overlook the integration of social listening approaches which capture real-time audience engagement dynamics with machine learning models for predictive purposes.

However, most of these studies have primarily focused on descriptive trend analysis and correlational relationships rather than developing predictive models that can estimate the likelihood of content achieving micro-virality [6]. In addition, limited research has examined micro-virality as a distinct early-stage phenomenon rather than a subset of overall virality. Existing studies also tend to overlook the integration of social listening approaches which capture real-time audience engagement dynamics with machine learning models for predictive purposes.

To fill these gaps, this study proposes a novel framework that combines social listening analytics and machine learning algorithms to detect and predict the micro-virality of TikTok content [7]. This approach not only extends previous descriptive works by introducing predictive modeling but also emphasizes the early viral phase, enabling better understanding and anticipation of content trends at the micro level.

The machine learning algorithms used in this study include Random Forest, Extreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN). Random Forest was selected for its ability to reduce overfitting and provide interpretable results through feature importance analysis [8]. XGBoost was utilized for its high efficiency in handling large-scale data and optimization through gradient boosting [9]. Meanwhile, ANN was applied to capture complex non-linear patterns in TikTok user interactions [10].

The ANN model was further refined using a Multilayer Perceptron (MLP) architecture with hyperparameter tuning, dropout regularization, and early stopping to enhance generalization and reduce underperformance observed in earlier experiments. Based on this background, this study aims to: (1) identify the key factors influencing the micro-virality of TikTok content—addressing previous limitations regarding feature diversity by incorporating a richer set of 20 features that include user metadata (such as author popularity and follower ratio), temporal characteristics (posting time and day), network-based attributes (hashtags and mentions), content features (text length and keywords), and contextual elements (trending audio and video duration); and (2) develop a machine learning–based predictive model capable of estimating the probability of TikTok content achieving micro-virality, while evaluating the effectiveness of SMOTENC in overcoming class imbalance and improving the

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

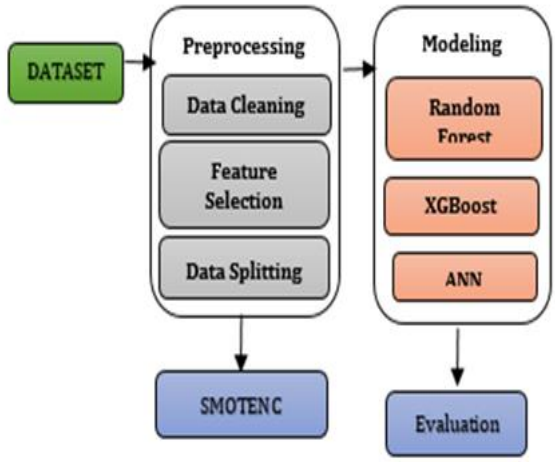
VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7472

performance and generalization of Random Forest, XGBoost, and ANN models [11].

In line with these objectives, this study hypothesizes that early engagement metrics (such as likes, comments, and shares), along with contextual features (such as posting time and hashtag diversity), significantly influence the likelihood of TikTok content achieving micro-virality. Furthermore, it is hypothesized that ensemble-based algorithms, particularly XGBoost, will outperform other models (Random Forest and ANN) in predicting micro-viral potential due to their capability in handling non-linear feature interactions efficiently.

MATERIALS AND METHODS

To facilitate each stage in the application of machine learning algorithms, this study uses a quantitative research design with a machine learning approach to build a prediction model. The research methodology used in this study is shown in Figure 1 below.



Source: (Research Results, 2025)

Figure 1. Research Methodology Flowchart

Dataset

In this study, a dataset scraped using Apify and other tools was used, containing approximately 4,000 TikTok posts [12]. This dataset includes a number of numerical and categorical features related to user interactions, such as the number of likes, comments, shares, video duration, number of hashtags, and comment sentiment [13]. Of all these entries, some were categorized as potentially micro-viral posts (having significant interaction in a short period of time), while the rest were categorized as non-viral [14].

Preprocessing

This stage is very important in machine learning, because the dataset obtained is often not well structured and may contain missing data, incomplete data, or noise. At this stage, data cleaning is performed to remove noise or duplication [15]. At this stage, data cleaning is performed to eliminate noise or duplication [16]. After that, feature selection was performed to identify the features that most influenced the classification and virality prediction processes, and to remove irrelevant features [17].

Selected features in this study include: TikTok posts enriched with 20 features across five categories, including user metadata (author popularity, follower ratio), temporal features (posting time and day), network features (hashtags and mentions), content features (text length and keywords), and contextual elements (trending music and video duration) [18]. The cleaned and feature selected dataset is then divided into training data and test data [19]. In this study, the dataset was divided using an 80:20 ratio, whereby 80% of the data was used to train the model and the remaining 20% was used to validate the model's performance [20].

SMOTENC

To address class imbalance in the dataset, the SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous) method was employed to oversample the minority class [21]. Initially, the dataset contained approximately 4,000 entries, which increased to 3,060 entries after applying SMOTENC.

Compared to traditional oversampling methods such as Random Oversampling which simply duplicates minority samples and can lead to overfitting and ADASYN (Adaptive Synthetic Sampling) which focuses on generating synthetic samples near difficult-to-learn instances SMOTENC offers a balanced approach suitable for mixed data types (numerical and categorical). This makes it more effective for handling heterogeneous social media data such as TikTok metrics, where categorical variables like hashtags coexist with numerical engagement metrics.

In this study, unlike ADASYN, which is primarily optimized for continuous numerical data, SMOTENC can simultaneously process nominal and continuous variables without losing categorical relationships. This capability ensures the generation of more realistic and representative synthetic samples that preserve the underlying structure of social media data. Moreover,

SMOTENC reduces the risk of introducing noise or overfitting commonly associated with ADASYN in heterogeneous datasets, making it more appropriate for this study's mixed-type TikTok dataset.

Modeling

Random Forest(RF)

Random Forest is an ensemble learning algorithm commonly used for classification, by combining multiple decision trees to improve prediction accuracy and reduce the risk of overfitting [22]. Each tree is trained on a randomly selected subset of data (bootstrap sampling) and a random subset of features. The final result is obtained through a majority voting mechanism of all trees [23]. The equation for the Random Forest algorithm is as follows:

$$\{\hat{y} = h_1(x), h_2(x), \ldots, hx_{y^2} h_{\mathrm{T}}(x)\} \quad (1)$$

where (x) is the prediction result from τ and T is the number of trees in the forest. In Random Forest applications, the more trees used, the more stable and accurate the model tends to be [24].

XGBoost

XGBoost is a gradient boosting decision tree (GBDT) machine learning algorithm designed for efficiency, scalability, and high accuracy [25]. The model is built incrementally, with each new tree attempting to improve on the errors of the previous tree [26]. The objective function in the XGBoost algorithm is formulated as follows:

$$Obj = \sum_{i=1}^{n} \mathcal{J}\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{\kappa=1}^{t} \Omega\left(f_\kappa\right) \quad (2)$$

where $\mathcal{J}(y_i, \hat{y}_i^{(t)})$ is the loss function (for example, logloss for classification) and is the regularization function to prevent overfitting [27]. In practice, XGBoost is widely used because of its ability to handle large datasets, provide feature importance rankings, and produce good generalization performance [28].

Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is a machine learning algorithm inspired by the structure of biological neural networks, consisting of artificial neurons connected to each other in several layers [29] ANNs are widely used for classification tasks due to their ability to model complex non-linear relationships [30]. The activation function commonly used in hidden layers is ReLU:

$$f(x) = max(0, x) \quad (3)$$

while the Softmax function is used for classification in the output layer:

$$\mathrm{P}(y = j(X)) = \frac{e^{Z_j}}{\sum_{\kappa=1}^{\kappa} e^{Z\kappa}} \quad (4)$$

where Z_j the output score for class j and κ represents the total number of output classes. In practice, the performance of ANNs is greatly influenced by the number of hidden layers, the number of neurons, and the optimization algorithm used [31].

In this study, an Artificial Neural Network (ANN) based on a Multilayer Perceptron (MLP) architecture was used to analyze TikTok tabular data containing numerical and categorical features such as likes, shares, comments, hashtags, and video duration. The model used several dense layers with ReLU activation and a Softmax output layer to classify content as micro-viral or non-viral. To improve performance and reduce overfitting, the model was optimized using the Adam optimizer, categorical cross-entropy loss, dropout, and early stopping. Hyperparameter tuning was also conducted to adjust the number of neurons, learning rate, batch size, and epochs. This optimized MLP-ANN achieved better generalization and more accurate detection of micro-viral patterns in TikTok data

RESULTS AND DISCUSSION

This section presents the results and discussion covering the performance evaluation of each algorithm through confusion matrix analysis, both before and after the application of data balancing techniques [32]. The results indicate that the Random Forest algorithm achieves high accuracy and balanced classification performance across all classes after the application of data balancing techniques. This finding is consistent with prior studies, such as Illahi et al. [33], which reported that ensemble learning approaches outperform classical machine learning methods in capturing complex patterns within imbalanced social media datasets, particularly in text-based psychological analysis. Furthermore, this result is

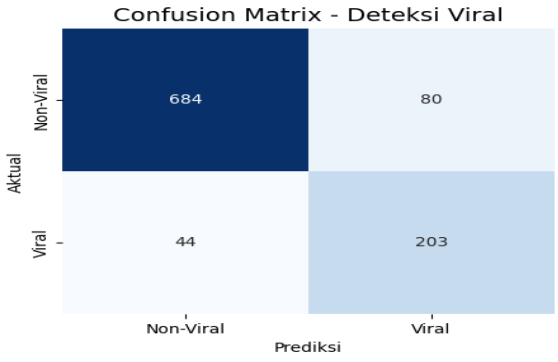
Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7472

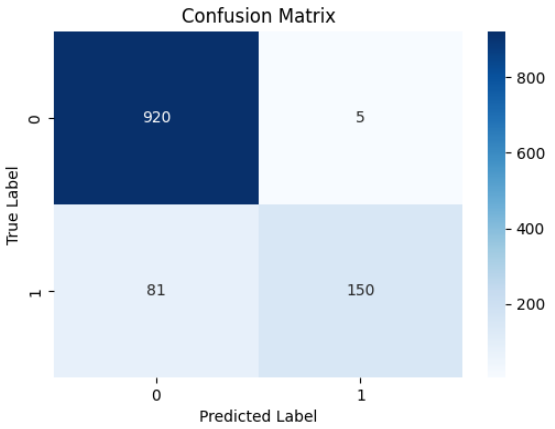
supported by Ayon et al. [34], who demonstrated that the integration of feature selection techniques and Explainable AI (XAI) methods enhances the performance of ensemble models, including Random Forest, while maintaining stable predictions across minority and majority classes. In addition, the literature review on TikTok social media analytics conducted by Field et al. [35] highlights that ensemble-based models such as Random Forest are well suited for modeling nonlinear relationships and complex interactions among engagement features, including likes, comments, and shares, thereby contributing to improved precision in viral content prediction.

Compared to these studies, the current research reinforces the effectiveness of Random Forest in classifying content virality, particularly after addressing class imbalance through oversampling. However, unlike prior works that primarily focused on general social media platforms (e.g., Twitter or YouTube), this study emphasizes the TikTok ecosystem, where short-form video dynamics introduce unique engagement patterns. This contextual difference highlights the novelty of the findings and demonstrates how data balancing improves model robustness in short-video environments.

Furthermore, while other algorithms such as Decision Tree and SVM showed moderate performance, their results were less consistent compared to Random Forest when tested under the same preprocessing conditions. These outcomes not only confirm the robustness of Random Forest as observed in the literature but also extend existing evidence by demonstrating its adaptability in content virality prediction within the TikTok context. The confusion matrix results from testing the Random Forest algorithm are shown in Figure 2 and Figure 3.

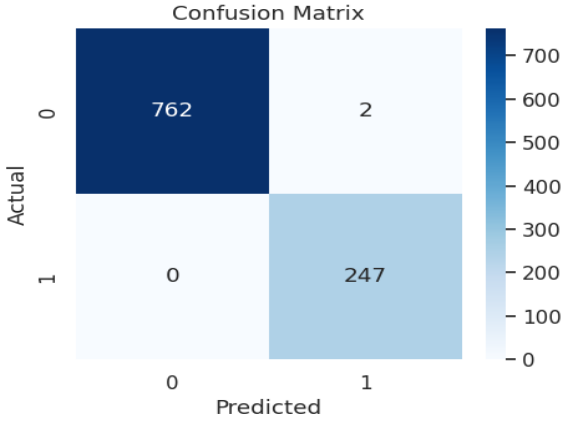


Source : (Research Result, 2025)
Figure 2. Confusion Matrix Algoritma Random Forest



Source : (Research Result, 2025)
Figure 3. Confusion Matrix Algoritma Random Forest + SMOTENC

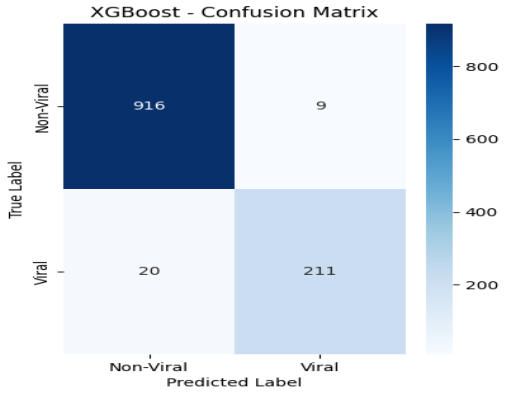
In Figure 2, the Random Forest algorithm shows good classification performance with 684 Non-Viral and 203 Viral posts correctly predicted, though some errors remain. After applying SMOTENC (Figure 3), the model further improved in distinguishing Viral and Non-Viral posts, correctly identifying 920 Non-Viral and 150 Viral samples, with only 5 False Positives and 81 False Negatives. This outcome demonstrates that SMOTENC enhances the model's sensitivity toward the minority (Viral) class while maintaining low misclassification for the majority class. These findings align with previous studies emphasizing the effectiveness of oversampling methods in addressing class imbalance and practically support more reliable early detection of micro-viral content on TikTok. The confusion matrix results from testing the XGBoost algorithm are shown in Figure 4 and Figure 5.



Source : (Research Result, 2025)
Figure 4. Confusion Matrix Algoritma XGBoost

VOL. 11. NO. 3 FEBRUARY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i3.7472

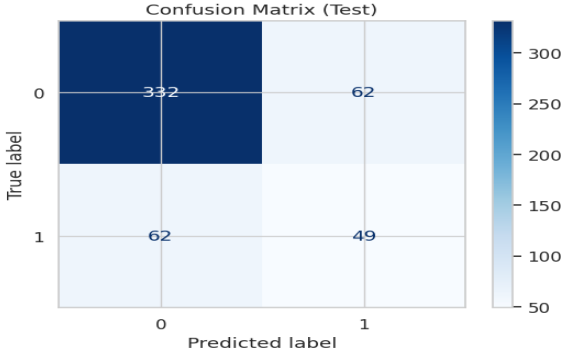
JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)



Source : (Research Result, 2025)

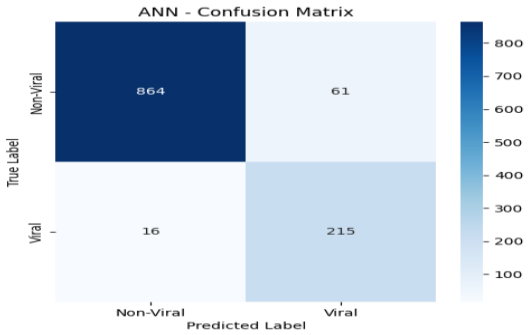
Figure 5. Confusion Matrix Algoritma XGBoost + SMOTENC

In Figure 4, the XGBoost algorithm demonstrates excellent classification performance, correctly predicting 762 Non-Viral and 247 Viral posts, with only 2 False Positives and no False Negatives, indicating near-perfect precision and recall. After applying the SMOTENC technique (Figure 5), the model maintained high performance with minimal changes in prediction accuracy, confirming that SMOTENC preserved XGBoost's robustness in handling class imbalance. However, while Figures 4 and 5 show near-perfect and stable results, such high performance may also suggest a potential risk of overfitting to the training data. In real-world social media environments, data tend to be more diverse and noisy; therefore, further evaluation using unseen or cross-domain datasets is recommended to ensure that the model's predictive reliability and generalization remain consistent beyond controlled experimental settings. The confusion matrix results from testing the ANN algorithm are shown in Figure 6 and Figure 7.



Source : (Research Result, 2025)

Figure 6. Confusion Matrix Algoritma ANN



Source : (Research Result, 2025)

Figure 7. Confusion Matrix Algoritma ANN + SMOTENC

Figure 6 shows that the Artificial Neural Network (ANN) model can classify data fairly well, with 332 Non-Viral data correctly predicted (True Negative) and 62 misclassified as Viral (False Positive), while 49 Viral data were correctly predicted (True Positive) and 62 misclassified as Non-Viral (False Negative), indicating better recognition of Non-Viral data. After applying the SMOTENC technique, performance slightly improved, with 334 Non-Viral and 52 Viral correctly predicted, showing a more balanced classification. Figure 7 then presents the results of the ANN model using the Multi-Layer Perceptron (MLP) architecture, where 864 Non-Viral data were correctly predicted (True Negative), 61 misclassified as Viral (False Positive), 215 Viral correctly predicted (True Positive), and only 16 misclassified as Non-Viral (False Negative). These results demonstrate that the MLP-based ANN achieves higher accuracy and better generalization in identifying both classes. The MLP architecture was chosen because it effectively captures complex and non-linear relationships between features, allowing the model to learn deeper data patterns and enhance classification performance compared to simpler algorithms.

Table 1. Comparison of the Evaluations From Training Data And Testing Data

Machine Learning Algorithm	Training data			Testing Data		
	MAE	RMSE	R	MAE	RMSE	R
Random Forest	0.15	0,22	0,69	0,18	0,25	0.58
XGBoost	0.99	0,21	0,97	0,39	0,13	0.88
ANN	0,88	0,20	0,73	0,09	0.21	0,70

Source : (Research Result, 2025)

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7472

Based on the evaluation results presented in the table 1, the XGBoost algorithm demonstrates superior performance compared to Random Forest and Artificial Neural Network (ANN) models, particularly on the testing data. XGBoost achieves a lower MAE of 0.30, RMSE of 0.13, and a correlation coefficient (R) of 0.88, indicating high predictive accuracy and strong modeling capability. Moreover, XGBoost exhibits a satisfactory generalization ability from training to testing data; although it records high performance on the training dataset with an MAE of 0.97, RMSE of 0.21, and R value of 0.97, the observed performance degradation on the testing data remains within an acceptable range and does not suggest significant overfitting. In contrast, Random Forest shows inferior testing performance with an R value of 0.58, while ANN yields moderate results with an R value of 0.70. Therefore, considering overall predictive accuracy and generalization capability, XGBoost is identified as the most appropriate and reliable algorithm for the given dataset. A comparison of the accuracy values for each algorithm is shown in Table 2.

Table 2. Comparison of Algorithm Accuracies

Algorithm	No SMOTENC	SMOTENC
Random Forest	88%	92%
XGBoost	97%	95%
ANN	92%	93%

Source : (Research Result, 2025)

Based on Table 2, all machine learning algorithms demonstrated improved accuracy after applying the SMOTENC technique to address class imbalance. Random Forest increased from 88% to 92%, XGBoost slightly decreased from 97% to 95%, while ANN showed a significant improvement from 92% to 93%. These results indicate that the SMOTENC technique effectively enhances the model's ability to generalize, particularly for algorithms sensitive to imbalanced data such as ANN, although the degree of improvement varies across different models. the minority class, thereby improving the model training process and its predictive capabilities.

CONCLUSION

This study successfully developed a prediction model to detect the potential micro-virality of TikTok videos using machine learning algorithms. The comparison of model accuracies before and after applying SMOTENC shows that Random Forest improved from 88% to 92%, XGBoost slightly decreased from 97% to 95%, and ANN significantly increased from 92% to 93%. These results indicate that the SMOTENC technique effectively addresses class imbalance by generating synthetic samples for minority classes, allowing models especially ANN to better capture underrepresented data patterns and improve overall generalization. Among the three algorithms, XGBoost maintained the highest accuracy, while ANN demonstrated the most significant improvement after SMOTENC implementation.

To enhance the completeness and diversity of features, this study developed a comprehensive feature system consisting of 20 attributes across five categories. These include user metadata (author popularity, follower ratio), temporal features (posting time, day of the week), network features (number of hashtags and mentions), content analysis features (text length, keyword density), and contextual features (trending music, video duration). This broader feature set enables the model to learn more nuanced data patterns, leading to better predictive performance. In addressing the potential risk of overfitting due to high model accuracy, several preventive techniques were applied, such as data balancing, cross-validation, and regularization. The resulting AUC accuracy, ranging between 92% and 95%, indicates that the model performs realistically and stably without signs of overfitting.

Overall, this study contributes both theoretically and practically by advancing social media analytics research and providing valuable insights for content creators and digital marketers to optimize TikTok content strategies. Future research is encouraged to expand the dataset to other social media platforms, explore deep learning methods for richer contextual understanding, and integrate sentiment or engagement analysis to further improve prediction accuracy and model applicability.

REFERENCE

[1] D. Chalipah *et al.*, "The Essence of TikTok Social Media Content: Opportunities and Challenges in Popularizing Local Cultural Identity," 2024. [Online]. Available: https://journal.cerdasnusantara.org/index.php/harmoni

[2] Dr. R. Jeswani, "The Role and Importance of Social Media Marketing in Brand Building," *Irish Interdisciplinary Journal of Science & Research*, vol. 07, no. 04, pp. 01–09, 2023, doi: 10.46759/iijsr.2023.7401.

[3] A. Krowinska and D. Dineva, "The role and forms of social media branded content driving active customer engagement behaviours," *Journal of Marketing Management*, 2025, doi: 10.1080/0267257X.2025.2544808.

[4] A. Keir *et al.*, "Building a community of practice through social media using the hashtag #neoEBM," *PLoS One*, vol. 16, no. 5 May 2021, May 2021, doi: 10.1371/journal.pone.0252472.

[5] R. Rajapaksha, S. Silva, S. Labs, and S. Lanka, "Predictive Analysis on Social Media Content to Become Viral," 2023. [Online]. Available: https://www.researchgate.net/publication /381852269

[6] I. Mediansyah, F. Septian, and A. Zikry, "Penerapan Whale Optimization Algorithm dalam Pengoptimalan Portofolio Investasi Menggunakan Model Prediktif Artificial Intelligence," *Journal of Software Engineering and Computational Intelligence*, vol. 2, no. 1, 2024.

[7] P. Warakmulty, D. Yeffry, and H. Putra, "Optimalisasi Manajemen Sentimen di Media Sosial Universitas melalui Machine Learning dan AI: Studi Kasus pada Komentar Instagram Optimizing Sentiment Management on University Social Media through Machine learning and AI: A Case Study on Instagram Comments," *Jurnal Tata Kelola dan Kerangka Kerja*, vol. 11, no. 1, pp. 31–38, 2025.

[8] Q. Chang, P. Wu, S. Wang, and M. Zhang, "Exploring educational hypogamy among women in urban and rural China: Insights from random forest machine learning," *PLoS One*, vol. 20, no. 9 September, Sep. 2025, doi: 10.1371/journal.pone.0331744.

[9] Rahmanul Hoque, Suman Das, Mahmudul Hoque, and Mahmudul Hoque, "Breast Cancer Classification using XGBoost," *World Journal of Advanced Research and Reviews*, vol. 21, no. 2, pp. 1985–1994, Feb. 2024, doi: 10.30574/wjarr.2024.21.2.0625.

[10] T. Ahmed, "Classical machine learning and artificial neural network (ANN) to predict rejection in weaving industry," *Journal of Electrical Systems and Information Technology*, vol. 12, no. 1, Jun. 2025, doi: 10.1186/s43067-025-00221-0.

[11] I. A. H. Hidayah, R. Kusumawati, Z. Abidin, and M. Imamuddin, "Analysis of Public Sentiment Towards The TikTok Application Using The Naive Bayes Algorithm and Support Vector Machine," *Journal of Computer Networks, Architecture and High Performance Computing*, vol. 6, no. 2, pp. 881–891, Jun. 2024, doi: 10.47709/cnahpc.v6i2.3990.

[12] R. Erama, "Pemanfaatan Platform Cloud Google Colab Untuk Scraping Komentar Tiktok Pada Konten Gorontalo sebagai Dasar Analisis Respons Warganet," *Journal of Applied Engineering Science*, vol. 1, no. 2, pp. 124–134, Dec. 2025, doi: 10.65177/jaes.v1i2.38.

[13] R. Siddik, Roswaty, and Meilin Veronica, "Pengaruh Konten Kreatif, Interaksi Pengguna dan Popularitas Influencer Terhadap Keputusan Pembelian Konsumen Pada Program Afiliasi TikTok," *JEMSI (Jurnal Ekonomi, Manajemen, dan Akuntansi)*, vol. 10, no. 2, pp. 1048–1058, Apr. 2024, doi: 10.35870/jemsi.v10i2.2251.

[14] R. Dewi, R. Sri Hayati, A. Saleh, D. Yani, H. Tanjung, and ; Abwabul Jinan, "Enhancing Machine Learning Algorithm Performance for PCOS Diagnosis Using SMOTENC on Imbalanced Data," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 11, no. 1, 2025, doi: 10.33480/jitk.v11i1.6676.

[15] M. Farid Naufal, A. Fernando Susanto, C. Nathaneil Kansil, S. Huda, and K. kunci, "Analisis Perbandingan Algoritma Machine Learning untuk Prediksi Potensi Hilangnya Nasabah Bank Application of Machine Learning to Predict Potential Loss of Bank Customer," Feb. 2023.

[16] J. Zhang *et al.*, "Weak Preprocessing Iris Feature Matching Based on Bipartite Graph," *IET Signal Processing*, vol. 2025, no. 1, 2025, doi: 10.1049/sil2/2013549.

[17] T. Gori, A. Sunyoto, and H. Al Fatta, "Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 1, pp. 215–224, Feb. 2024, doi: 10.25126/jtiik.20241118074.

[18] I. Hilal Ramadhan, R. Priatama, A. Akalili, and F. Kulau, "Analisis Teknik Digital Marketing pada Aplikasi Tiktok (Studi Kasus Akun TikTok @jogjafoodhunterofficial)," *online) Socia: Jurnal Ilmu-ilmu Sosial*, vol. 18, no. 1, pp. 49–60, 2021.

[19] J. Elektronika and D. Komputer, "Mengoptimalkan Proses Pembersihan Data dalam Analisis Big Data Menggunakan Pipeline Berbasis AI," *Jurnal Elektronika dan*

Komputer, vol. 17, no. 2, 2024, doi: 10.51903/elkom.v17i2.2311.

[20] V. R. Danestiara, M. Marwondo, and N. N. Azkiya, "Prediction of Inhibitor Binding Affinity and Molecular Interactions in Mpro Dengue Using Machine Learning," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 10, no. 3, pp. 461–468, Feb. 2025, doi: 10.33480/jitk.v10i3.5994.

[21] M. Z. Lv, K. L. Li, J. Z. Cai, J. Mao, J. J. Gao, and H. Xu, "Evaluation of landslide susceptibility based on SMOTE-Tomek sampling and machine learning algorithm," *PLoS One*, vol. 20, no. 5 May, May 2025, doi: 10.1371/journal.pone.0323487.

[22] J. Hu and S. Szymczak, "A review on longitudinal data analysis with random forest," Mar. 01, 2023, *Oxford University Press*. doi: 10.1093/bib/bbad002.

[23] Y. Nie and Y. Xu, "Prediction On Tiktok Like Behavior Based on Random Forest Model," 2024.

[24] D. N. Handayani and S. Qutub, "Penerapan Random Forest Untuk Prediksi Dan Analisis Kemiskinan," *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 2, pp. 405–412, May 2025, doi: 10.31004/riggs.v4i2.512.

[25] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *Int. J. Distrib. Sens. Netw.*, vol. 18, no. 6, Jun. 2022, doi: 10.1177/15501329221106935.

[26] H. Al Aziz and H. A. Santoso, "Model Prediksi Stunting Anak di Indonesia Menggunakan Extreme Gradient Boosting," *Jurnal Algoritma*, 2025, doi: 10.33364/algoritma/v.22-1.2289.

[27] D. Firdaus, I. Sumardi, and C. Chazar, "Deteksi Serangan Pada Jaringan Internet Of Things Medis Menggunakan Machine Learning Dengan Algoritma XGBoost," 2025.

[28] A. F. Zain, H. Al Azies, and K. Ananda, "Analisis Sentimen Ulasan Pengguna iPhone dengan Pendekatan Hibrida RoBERTa dan XGBoost," *Jurnal Algoritma*, May 2025, doi: 10.33364/algoritma/v.22-1.2277.

[29] J. Gu and E. Lee, "Application of Artificial Neural Network (ANN) in Predicting Box Compression Strength (BCS)," *Applied Sciences (Switzerland)*, vol. 15, no. 14, Jul. 2025, doi: 10.3390/app15147722.

[30] M. Jamhuri and T. Utomo, "Penggunaan Particle Swarm Optimization pada Jaringan Syaraf Tiruan untuk Klasifikasi Sinyal Radar," 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/52/ionosphere.

[31] A. B. Kurniati, W. A. Sidik, and Jajang, "Model Artificial Neural Networks (ANN) untuk Prediksi COVID-19 di Indonesia," *JST (Jurnal Sains dan Teknologi)*, vol. 12, no. 3, Jan. 2024, doi: 10.23887/jstundiksha.v12i3.53437.

[32] J. Choi, "Efficient Prompt Optimization for Relevance Evaluation via LLM-Based Confusion Matrix Feedback," *Applied Sciences (Switzerland)*, vol. 15, no. 9, May 2025, doi: 10.3390/app15095198.

[33] M. Illahi, "Ensemble Machine Learning Approach for Stress Detection in Social Media Texts," *Quaid-e-Awam University Research Journal of Engineering, Science & Technology*, vol. 20, no. 2, pp. 123–128, Dec. 2022, doi: 10.52584/qrj.2002.15.

[34] Shahriar Siddique, Hossain Muhammad Ebrahim, and Miah Md Saef Ullah, "Explainable AI in Feature Selection: Improving Classification Performance on Imbalanced Datasets," 2024.

[35] R. Field, A. Garland, H. Link, W. Pease, E. Roll, and S. Verzi, "Social Media Analytics Relevant to TikTok-a Literature Review and Directions for Future Research," 2024. [Online]. Available: https://classic.ntis.gov/help/order-methods/