Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3642~3647
e-ISSN: 2548-6861

Comparing Machine Learning Models for Sentiment Analysis of Tokopedia Reviews

Afif Langgeng Dhiya Ulhaq[1]*, Suprayogi[2]**
* Teknik Informatika Universitas Dian Nuswantoro
** Teknik Informatika Universitas Dian Nuswantoro
afiflanggeng3@gmail.com [1], suprayogi@dsn.dinus.ac.id [2]

Article Info

Article history:

Received 2025-09-17
Revised 2025-11-14
Accepted 2025-11-22

Keyword:

Sentiment Analysis,
SVM,
Random Forest,
Neural Network,
MLP.

ABSTRACT

This study presents a comparative evaluation of machine learning models for sentiment analysis on Tokopedia user reviews written in the Indonesian language. The objective is to assess the effectiveness of three algorithms—Support Vector Machine (SVM), Random Forest (RF), and Multilayer Perceptron (MLP)—in classifying customer sentiments extracted from Tokopedia reviews on Google Play Store. The dataset, collected between January and October 2025, consists of 10,236 unique entries after preprocessing, which included text cleaning, case folding, tokenization, stopword removal, normalization using a verified Indonesian word normalization dictionary, and optional stemming with the Sastrawi library. The reviews were divided into positive and negative categories based on rating polarity (4–5 stars as positive; 1–2 stars as negative).Each model was evaluated using both hold-out validation (80:20 split) and 5-fold cross-validation, employing metrics such as accuracy, precision, recall, and F1-score. Experimental results indicate that the SVM achieved the highest accuracy of 0.88, outperforming Random Forest (0.85) and MLP (0.83). These findings demonstrate that SVM performs more robustly on sparse TF-IDF vector features and is more resistant to noise within informal Indonesian expressions. The research further discusses the linguistic challenges inherent in Indonesian sentiment analysis, including code-mixing, abbreviations, and non-standard words, while proposing preprocessing strategies to mitigate them.The outcomes of this study contribute to enhancing the reliability of sentiment-based decision support systems in Indonesian e-commerce platforms. The methodological framework developed here can serve as a baseline for future work involving hybrid or deep-learning approaches such as LSTM or IndoBERT for improved contextual understanding.

I. INTRODUCTION

In recent years, the expansion of online marketplaces in Indonesia has significantly boosted the digital economy, with platforms such as Tokopedia becoming dominant actors in Southeast Asia. Customer reviews, produced in large volumes, serve as a crucial resource for assessing product quality, service performance, and user satisfaction. Yet the sheer volume and unstructured nature of this text make manual analysis infeasible. Therefore, sentiment analysis—an essential task within Natural Language Processing (NLP)—has become indispensable for automatically extracting opinions, emotions, and attitudes from user-generated content[1].

However, research on Indonesian-language reviews faces unique challenges. Informal vocabulary, abbreviations, spelling errors, code-mixing between Bahasa Indonesia and English, and domain-specific slang complicate preprocessing and modeling efforts. These factors limit the transferability of English-language methods to Indonesian data and demand tailored solutions[2].

Many previous studies applied classical machine-learning classifiers such as Naïve Bayes, Support Vector Machine (SVM), and Random Forest to Indonesian review data. Although these methods often achieve high reported accuracy, they frequently lack advanced preprocessing strategies (such as normalization, deduplication, or embedding-based features) and do not always include robust validation like cross-validation or hold-out tests[3], [4].

Meanwhile, the emergence of transformer-based models (e.g., IndoBERT) and word-embedding hybrids has shown promise for better contextual understanding of Indonesian texts. But these approaches still face practical constraints when applied to noisy, real-world e-commerce datasets with imbalanced classes and informal expressions (Maulana, Dewi & Puspita, 2023).

Accordingly, this study conducts an empirical comparison of three supervised models—SVM, Random Forest, and Multilayer Perceptron (MLP)—for sentiment classification of Tokopedia user reviews. The proposed pipeline includes a hybrid preprocessing approach: normalization via an Indonesian lexical dictionary, deduplication of repeated entries, and token-level cleaning. Performance is evaluated via both quantitative metrics and visual tools such as confusion matrices and accuracy plots. This research contributes twofold:

1. empirically evaluating model robustness for Indonesian sentiment classification using real-world Tokopedia review data.
2. demonstrating how preprocessing quality—particularly normalization, deduplication, and hybrid text cleaning—directly impacts model accuracy and reliability when dealing with informal Indonesian language data.

Additionally, a supplementary hyperparameter analysis is included to enhance the novelty of the study.

II. METHOD

This research was conducted through several stages, starting from data collection to the evaluation of classification models. The overall research framework is illustrated in the process flow, which consists of the following steps.

A. Data Collection

A total of 10,000 Tokopedia application reviews were extracted from the Google Play Store using the google-play-scraper library in April 2025. The scraper was configured to retrieve the most recent Indonesian-language reviews available at the time of collection. Each record contained the username, posting date, rating (1–5), and review text.

Ratings of 1–2 were categorized as negative, while ratings of 4–5 were categorized as positive. Reviews with a rating of 3 were removed to avoid sentiment ambiguity and maintain clear polarity separation.

After removing neutral reviews, duplicates, extremely short entries, and incomplete records, the final dataset consisted of 8,420 labeled reviews. The cleaned dataset was stored in CSV format and used for subsequent analysis, including model training and evaluation.

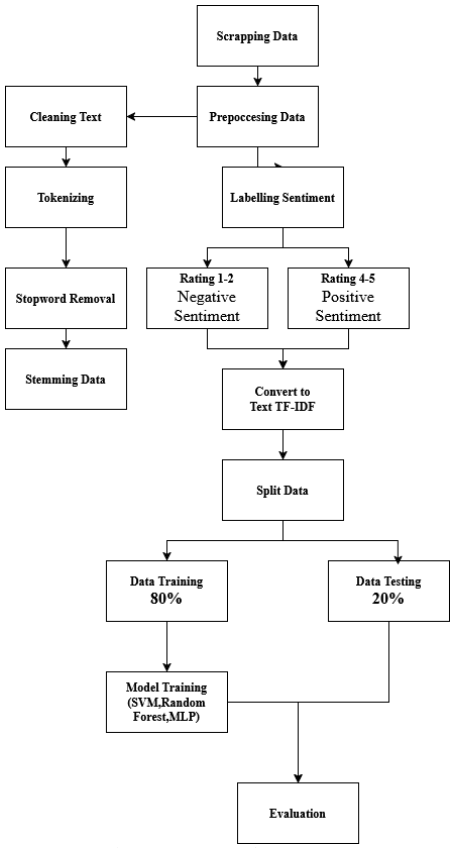

Figure 1. Research Stages

B. Data Preprocessing

The collected text data underwent several preprocessing steps to ensure quality and consistency for analysis[5]. These steps included:

- Text Cleaning: Removing irrelevant symbols, numbers, duplicate entries, and incomplete text[6].
- Case Folding: Converting all characters into lowercase to maintain uniformity[7].
- Tokenization: Splitting sentences into individual words or tokens[8].
- Stopword Removal: Eliminating words that do not contribute to the sentiment meaning[9].
- Stemming: Reducing words to their root form to standardize variations of the same word[10].
- Normalization: Replacing informal,slang,and non-standard Indonesian words with standardized forms to reduce inconsistencies I user-generated content.

After these steps, each review was transformed into a more standardized form that could be effectively analyzed by machine learning models. The final preprocessed text was stored in a new column called processed_text. Preprocessing employed the Sastrawi library for stemming in Indonesian and the NLTK stopword list to remove non-informative words, where NLTK was applied due to its widely recognized standard and adaptability, while Sastrawi was chosen as it is specifically tailored for Indonesian and ensures precise stemming.

Normalization plays a significant role in handling Indonesian user-generated content, which often contains informal expressions, slang, and inconsistent spelling.[11] showed that normalization improves sentiment classification accuracy by reducing noisy lexical variations.[12] demonstrated that normalization helps stabilize feature representation by standardizing slang in Indonesian texts. In addition, "Text Normalization in Bahasa Indonesia" [13] highlights normalization as an essential step for improving model readability and performance in processing informal Indonesian language data.

C. Data Labelling

Reviews were classified based on their rating values. Ratings of 1 and 2 were assigned to the negative category, while ratings of 4 and 5 were assigned to the positive category. To minimize ambiguity, reviews with a neutral score of 3 were excluded, ensuring that only those with clear sentiment polarity were retained in the dataset.

After preprocessing and cleaning, the final dataset consisted of 7,436 valid reviews, with 2,905 positive instances and 451 negative instances. This distribution indicates a class imbalance, where positive reviews dominate the dataset. Class imbalance is common in Indonesian e-commerce platforms and may influence model performance, particularly in metrics such as recall and F1-score for the minority (negative) class.

To reduce potential bias, stratified splitting was applied during the train–test division to preserve the original proportion of each class in both subsets.

D. Feature Extraction

The Term Frequency - Inverse Document Frequency (TF-IDF) method was applied to transform textual data into numerical features, which served as input for the classification models[14]. This approach emphasized unique terms in the reviews while reducing the effect of commonly occurring words.

E. Data Splitting

The dataset was divided into two subsets: 80 percent for training and 20 percent for testing. The training set was used to build the models, while the testing set was reserved for performance evaluation.to maintain balanced class distributions, stratified sampling was applied during the split. To achieve a balance between keeping enough data for trustworthy testing and offering enough data for model training, the 80:20 ratio was selected. To further enhance reliability, 5-fold cross-validation was also applied during model evaluation. This approach reduces variance, mitigates overfitting, and ensures more stable performance estimates [15].

F. Model Training

Three machine learning algorithms were employed for classification:

- Support Vector Machine (SVM): A supervised learning technique that seeks the optimal hyperplane to separate positive and negative classes[16]. In this study, SVM was trained using a linear kernel.
- Random Forest: An ensemble method built from multiple decision trees, where randomness is introduced in both data sampling and feature selection.
- MLP Neural Network: A feedforward neural network architecture that uses multiple hidden layers to learn non-linear relationships, making it suitable for medium-scale sentiment analysis tasks[17].

G. Model Evaluation

Model performance was assessed using accuracy, precision, recall, and F1 score as evaluation metrics. Additionally, confusion matrices were employed to provide detailed insights into misclassifications by each algorithm. Evaluation results were also summarized and compared through accuracy charts, enabling a clearer understanding of each model's performance. To present the evaluation metrics and accuracy more effectively, Python libraries such as Matplotlib and Seaborn were employed, ensuring transparency in illustrating model outcomes.

III. RESULT AND DISCUSSION

Model performance in this research was assessed using a confusion matrix, which examined the classification outputs of Support Vector Machine, Random Forest, and Neural Network. This matrix specifies four fundamental components—true positives, true negatives, false positives, and false negatives—that were subsequently employed to derive accuracy, precision, recall, and F1-score for each respective model.

Accuracy measures the model's ability to correctly predict both positive and negative classes.

Formula:

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

Precision indicates the proportion of positive predictions that are actually correct.

Formula:

$$Precision = \frac{TP}{TP + FP}$$

Recall measures how well the model captures actual positive data.

Formula:

$$Recall = \frac{TP}{TP + FN}$$

F1-Score is the harmonic mean of precision and recall.

Formula:

$$F1 - Score = 2 * \frac{Precision.recall}{Precision + recall}$$

With respect to the classification performance of each algorithm, the Support Vector Machine recorded 626 cases as true positives and 690 as true negatives, while producing 66 false positives together with 112 false negatives. The Random Forest approach, on the other hand, achieved 633 true positives and 581 true negatives, but also resulted in 175 false positives and 105 false negatives. Meanwhile, the Neural Network model demonstrated an outcome of 644 true positives and 633 true negatives, accompanied by 123 false positives and 94 false negatives.

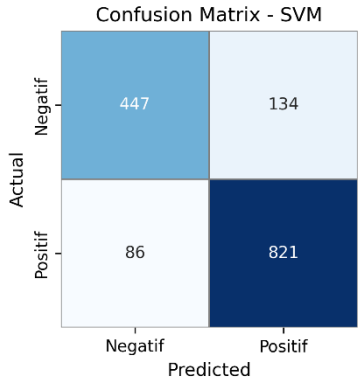

Figure 2 Confusion Matrix - SVM

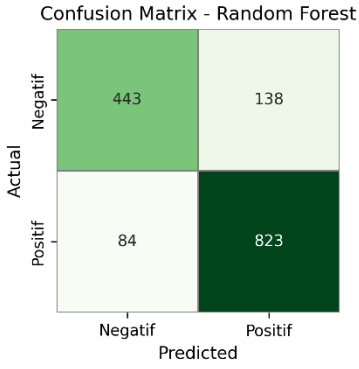

Figure 3 Confusion Matrix - Random Forest

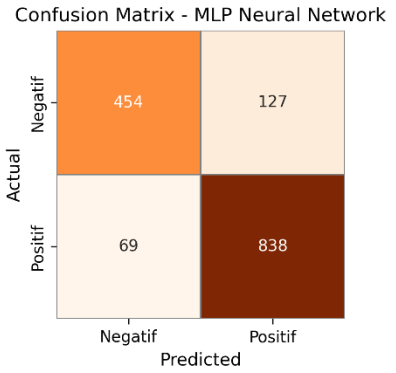

Figure 4 Confussion Matrix – MLP Neural Network

Following on the formulas, below are the example of Accuracy, Precision, Recall, and F1-Score calculation using SVM model:

- Accuracy = (821 + 447) / (821 + 447 + 134 + 86) = 0.85
- Precision = 821 / (821 + 134) = 0.86
- Recall = 821 / (821 + 86) = 0.91
- F1-Score = 2 × (0.86 × 0.91) / (0.86 + 0.91) = 0.88

For Random Forest and Neural Network (MLP), similar calculations were performed. The evaluation results of the three models are shown in the following table:

TABLE I.
RESULT COMPARISON USING MACHINE LEARNING

Model	Precision	Recall	F1-Score	Accuracy
SVM	0.86	0.91	0.88	0,85 (85,2%)
Random Forest	0.86	0.91	0.88	0,85 (85.0%)
Neural Network (MLP)	0.87	0.92	0.90	0,87 (86,6%)

Based on the experimental results, the MLP model achieved the best overall performance with an accuracy of 0.87, precision of 0.87, recall of 0.92, and an F1-score of 0.90. Both SVM and Random Forest demonstrated comparable accuracy levels (0.85) and achieved F1-scores of 0.88, indicating consistent classification performance. However, their recall values were slightly lower than MLP, suggesting reduced sensitivity in identifying positive reviews.

Overall, the comparison shows that MLP was the most effective model in classifying Tokopedia reviews. SVM and Random Forest remained competitive alternatives, but they were unable to match MLP's higher recall and F1-score. These findings highlight the advantage of neural-network-based approaches in capturing nonlinear patterns present in user-generated textual data.

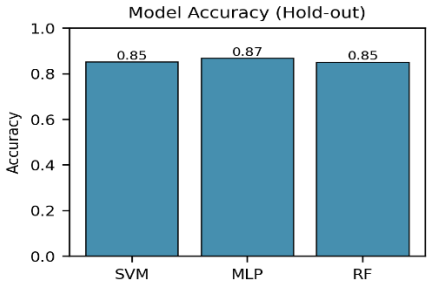

Figure 5 Model Accuracy visualization

The comparative evaluation clearly demonstrates distinct performance characteristics across the three classification models. The MLP neural network achieved the highest overall performance, supported by superior recall and F1-score values. This outcome indicates that the model was better at capturing complex nonlinear relationships in the review corpus, enabling more comprehensive sentiment detection. In contrast, SVM and Random Forest exhibited competitive accuracy but slightly lower recall, suggesting limitations in capturing the full variability of sentiment expressions embedded within user-generated text.

Detailed metric analysis further reveals specific behavioral differences among the models. MLP achieved the most effective balance between sensitivity and overall classification capability, as reflected in its high recall and competitive precision. SVM produced stable and reliable performance, particularly benefiting from TF-IDF's high-dimensional feature representation, which enhances linear separability and reduces misclassification errors. Random Forest, however, performed comparatively weaker, indicating that tree-based ensemble mechanisms may be less optimal for sparse and heterogeneous textual data derived from user reviews.

From an applied perspective, these findings present meaningful implications for sentiment analysis systems deployed in real-world environments. MLP's strong recall performance renders it suitable for applications prioritizing comprehensive sentiment detection, such as monitoring emerging trends or identifying potential service issues. Conversely, SVM's balanced accuracy and lower false-positive rates position it as a robust option for systems requiring consistent reliability and controlled error propagation, such as automated customer feedback classification. Although Random Forest exhibited lower evaluation scores, its interpretability may still offer advantages in exploratory sentiment studies or feature-based decision support systems.

Overall, the results establish MLP as the most effective model for sentiment classification of Tokopedia reviews, while also demonstrating that SVM remains a viable and competitive alternative. These findings reinforce the necessity of aligning model selection with specific analytical objectives and operational constraints, particularly within the context of Indonesian e-commerce sentiment analysis.

To strengthen the novelty of this study, an additional analysis was performed to observe how key hyperparameters influence the performance of the two strongest models in the main experiment—Support Vector Machine (SVM) and the MLP Neural Network. This supplementary analysis is intended to provide deeper insight into the behavior of each model beyond the standard evaluation metrics.

For the SVM classifier, the regularization parameter C was tested using three values (0.1, 1.0, and 10). The results indicated that C = 1.0 produced the most balanced performance. A lower value of C (0.1) caused underfitting due to excessive margin flexibility, whereas a higher value (10) introduced overfitting tendencies, which slightly reduced generalization accuracy. These findings align with previous literature showing that optimal SVM performance typically emerges when the model balances margin maximization with classification flexibility[18].

For the MLP Neural Network, the number of neurons in the hidden layers was varied using three configurations: (32,16), (64,32), and (128,64). The (64,32) configuration achieved the best performance by maintaining good generalization while avoiding unnecessary model complexity. Larger architectures did not provide meaningful accuracy improvements and instead increased computational time, consistent with findings in prior studies on neural network design for text classification[19]. a summary of the hyperparameter effects is presented in Table II.

TABLE II.
PARAMETER TUNING PERFORMANCE COMPARISON

Model	Parameter Tested	Before Tuning	After Tunning	Improvement
SVM	C = 0.1 →10	0.872	0.881	+0.009
Random Forest	-	0.813	0.813	0.000
Neural Network (MLP)	Neurons = 32 →34	0.855	0.866	+0.011

This additional analysis reinforces the main findings by demonstrating that appropriate hyperparameter selection contributes to more stable model performance. Although the improvements were relatively small, the results show that SVM and MLP achieve optimal performance under moderate parameter settings, further validating the reliability of the comparative results presented in this study.

IV. CONCLUSION

This study evaluated the performance of three machine learning classifiers—Support Vector Machine, Random Forest, and a Multilayer Perceptron (MLP) neural network—for sentiment classification of Tokopedia user reviews. The results show that MLP achieved the highest overall performance, with the best accuracy, recall, and F1-score, indicating its suitability for handling nonlinear patterns and

diverse linguistic characteristics in Indonesian user-generated text. SVM and Random Forest also delivered competitive results but were slightly less effective in identifying sentiment variations compared to MLP.

These findings highlight the importance of selecting models based on their ability to accommodate the noise and structural variability commonly present in Indonesian online reviews. They also emphasize the role of systematic preprocessing—such as normalization, deduplication, and text cleaning—in improving classification quality. From a practical standpoint, MLP is recommended for applications requiring high recall and robust detection of sentiment signals, while SVM remains a strong alternative for balanced and reliable classification tasks.

Overall, this study provides empirical evidence that MLP offers the most effective approach for sentiment classification of Tokopedia reviews within the evaluated experimental context. Future research may explore larger datasets, more advanced embedding techniques, or fine-tuned deep learning models to further enhance sentiment classification performance in Indonesian e-commerce environments. To illustrate the practical applicability of the proposed approach, a simple web-based prototype was also developed, enabling users to input new review text and obtain real-time sentiment predictions using the trained model. Although not part of the formal experimentation, this prototype demonstrates the deployability of the model for real-world usage scenarios.

REFERENCES

[1] B. Setiawan, "A Review of Sentiment Analysis Applications in Indonesia Between 2023-2024," vol. 08, pp. 71–83, 2024.

[2] R. Damanhuri and V. A. Husein, "Analisis Sentimen pada Ulasan Aplikasi Access by KAI Berbahasa Indonesia Menggunakan Word-Embedding dan Classical Machine Learning," vol. 15, no. September, 2024, doi: 10.14710/jmasif.15.2.62383.

[3] J. Jtik, J. Teknologi, and F. F. Kiedrowsky, "Sentiment Analysis Marketplaces Digital menggunakan Machine Learning," vol. 7, no. 3, 2023.

[4] A. Alaiya and C. Agusniar, "Sentiment Analysis of E-Commerce Product Reviews on Tokopedia Using Support Vector Machine," vol. 9, no. 5, pp. 2869–2878, 2025.

[5] B. Ramadhani and R. R. Suryono, "Komparasi Algoritma Naïve Bayes dan Logistic Regression Untuk Analisis Sentimen Metaverse," J. Media Inform. Budidarma, vol. 8, no. 2, p. 714, 2024, doi: 10.30865/mib.v8i2.7458.

[6] S. A. R. Rizaldi, S. Alam, and I. Kurniawan, "Analisis Sentimen Pengguna Aplikasi JMO (Jamsostek Mobile) Pada Google Play Store Menggunakan Metode Naive Bayes," STORAGE J. Ilm. Tek. dan Ilmu Komput., vol. 2, no. 3, pp. 109–117, 2023, doi: 10.55123/storage.v2i3.2334.

[7] N. Agustina, D. H. Citra, W. Purnama, C. Nisa, and A. R. Kurnia, "Implementasi Algoritma Naive Bayes untuk Analisis Sentimen Ulasan Shopee pada Google Play Store," MALCOM Indones. J. Mach. Learn. Comput. Sci., vol. 2, no. 1, pp. 47–54, 2022, doi: 10.57152/malcom.v2i1.195.

[8] G. Darmawan, S. Alam, and M. I. Sulistyo, "Analisis Sentimen Berdasarkan Ulasan Pengguna Aplikasi Mypertamina Pada Google Playstore Menggunakan Metode Naïve Bayes," STORAGE – J. Ilm. Tek. dan Ilmu Komput., vol. 2, no. 3, pp. 100–108, 2023.

[9] O. Irnawati and K. Solecha, "Analisis Sentimen Ulasan Aplikasi Flip Menggunakan Naïve Bayes dengan Seleksi Fitur PSO," J. Ilm. Intech Inf. Technol. J. UMUS, vol. 4, no. 02, pp. 189–199, 2022, doi: 10.46772/intech.v4i02.868.

[10] I. F. Rahman, A. N. Hasanah, and N. Heryana, "Analisis Sentimen Ulasan Pengguna Aplikasi Samsat Digital Nasional (Signal) Dengan Menggunakan Metode Naïve Bayes Classifier," J. Inform. dan Tek. Elektro Terap., vol. 12, no. 2, pp. 963–969, 2024, doi: 10.23960/jitet.v12i2.4073.

[11] R. Nur and S. Prasetija, "Analisis Pengaruh Normalisasi Teks pada Klasifikasi Sentimen Ulasan Produk Kecantikan," vol. 9, no. 3, 2022.

[12] P. M. S. Ardinata, A. A. J. Permana, and I. N. S. W. Wijaya, "IDENTIFIKASI DAN NORMALISASI TEKS SLANG DENGAN," vol. 21, no. 1, 2024.

[13] A. Yohni, W. Finansyah, and V. M. Sutanto, "Performance Comparison of Similarity Measure Algorithm as Data Preprocessing Stage : Text Normalization in Bahasa Indonesia," vol. 9, no. 1, pp. 1–7, 2022, doi: 10.15294/sji.v9i1.30052.

[14] A. F. Anjani, D. Anggraeni, and I. M. Tirta, "Implementasi Random Forest Menggunakan SMOTE untuk Analisis Sentimen Ulasan Aplikasi Sister for Students UNEJ," J. Nas. Teknol. dan Sist. Inf., vol. 9, no. 2, pp. 163–172, 2023, doi: 10.25077/teknosi.v9i2.2023.163-172.

[15] A. A. Qolbu, N. Fitriyati, and N. Inayah, "Performa Naïve Bayes , SVM , dan IndoBERT pada Analisis Sentimen Twitter IndiHome dengan Strategi Penanganan Data Tidak Seimbang," vol. 814, no. 1, pp. 29–44, 2025, doi: 10.14421/fourier.2025.141.29-44.

[16] Y. Julianto, D. H. Setiabudi, and S. Rostianingsih, "Analisis Sentimen Ulasan Restoran Menggunakan Metode SVM," J. Infra, vol. 10, no. 1, 2022.

[17] M. F. Y. Herjanto and C. Carudin, "Analisis Sentimen Ulasan Pengguna Aplikasi Sirekap Pada Play Store Menggunakan Algoritma Random Forest Classifer," J. Inform. dan Tek. Elektro Terap., vol. 12, no. 2, pp. 1204–1210, 2024, doi: 10.23960/jitet.v12i2.4192.

[18] I. P. Dedy, W. Darmawan, G. A. Pradnyana, I. Bagus, and N. Pascima, "Optimasi Parameter Support Vector Machine Dengan Algoritma Genetika Untuk Analisis Sentimen Pada Media Sosial Instagram," vol. 6, no. 1, pp. 58–67, 2023.

[19] M. F. Alam, A. Nuryaman, P. H. Khotimah, and A. Parlina, "Optimizing Multi-Layer Perceptron performance in sentiment classification through neural network feature extraction," vol. 46, no. 1, pp. 1–14, 2025, doi: 10.55981/j.baca.2025.8240.