International Journal of Health Science (IJHS)

E-ISSN: 2827-9603
P-ISSN: 2827-9603

Research Article

Comparison of Support Vector Machine and Random Forest Classification Methods in Dengue Fever Sentiment Analysis

Anisa Ranindia Arizky[1], Riska Fitria Ulandari[2], Grace Romauli Sihombing[3*], Nurul Baiah[4], Frandianus[5]

[1-5] Informatika Medis / Universitas Widya Husada Semarang
* Corresponding Author: graceromauli04@gmail.com

Abstract: Dengue fever is a prevalent disease in Indonesia, caused by the dengue virus and transmitted by Aedes aegypti and Aedes albopictus mosquitoes. This study aims to analyze sentiment in tweets related to dengue fever using two machine learning algorithms: Support Vector Machine (SVM) and Random Forest. The dataset consists of 1,000 manually labeled tweets, categorized into three sentiment classes: Positive, Negative, and Neutral. The results indicate that the SVM algorithm outperformed Random Forest in terms of performance metrics. Specifically, SVM achieved higher accuracy (77.07%) compared to Random Forest (73.65%). Additionally, SVM demonstrated superior precision, recall, and F1-score, making it a more effective model for sentiment analysis in this context. These findings suggest that SVM is a promising tool for analyzing public sentiment on social media platforms, particularly for monitoring health-related issues like dengue fever. The study highlights the potential of machine learning techniques in improving public health surveillance and response by analyzing social media data for real-time insights.

Keywords: DBD; Random; Sentiment; Support Vector Machine; Twitter.

1. Introduction

Dengue Hemorrhagic Fever (DHF) is a serious public health problem in tropical and subtropical countries, including Indonesia. Dengue fever is an infectious disease caused by the dengue virus. It is transmitted through the bite of the Aedes aegypti and Aedes albopictus mosquitoes, which act as vectors. Dengue fever can occur year-round and affects all age groups (I Gede Willy Karya Mahardika, 2023).

Dengue fever can affect anyone. Factors contributing to its occurrence include a dirty environment, prolonged rainy seasons, overcrowding, and the presence of stagnant water, which encourages the breeding of Aedes aegypti mosquitoes, leading to the emergence of mosquito larvae. Dengue fever can be prevented by maintaining environmental hygiene (Restu Wulandari et al., 2024).

In 2023, the WHO declared dengue fever a public health emergency following an increase in outbreaks in several countries. Several factors contribute to this increase, including very high mosquito populations, high rainfall, and humidity. According to estimates, there were 7.6 million cases in 2024, including 3.4 million confirmed cases, 16,000 of the most severe cases, and over 3,000 deaths. While cases have been reported globally in the past five years, the increase is particularly pronounced in the Americas, where the number of cases surpassed seven million by the end of April 2024, surpassing the annual record of 4.6 million cases in 2023. According to WHO data, the number of dengue fever cases reaches tens of millions annually, with a significant increase generally occurring during the rainy season (WHO, 2024).

Received: 14 February, 2026
Revised: 20 March, 2026
Accepted: 12 April, 2026
Online Available: 15 April, 2026
Curr. Ver.: 15 April, 2026

Despite various efforts, such as the 3M Plus Mosquito Nest Eradication Program (PSN), which includes draining or cleaning, covering, and reusing used items, these efforts remain a challenge for the community in preventing dengue fever (Ratna Dian K, 2022).

With the rapid development of technology, social media platforms like Twitter have opened up opportunities to monitor health issues through epidemiology, including data collection and analysis of dengue fever cases. Data generated from social media, including Twitter, can realistically reflect public perception and has been used in various studies to monitor various diseases (Prabaswara & Saputra, 2020). Twitter is a microbiology-based social media platform that allows users to express complaints, opinions, send short messages, and seek information related to health and social issues. One health issue currently under serious scrutiny is dengue fever (DHF).

Previous research by Prabaswara and Saputra (2020) found a significant correlation between DHF cases on social media and official data reported by health agencies. Therefore, social media-based sentiment analysis, or Twitter, is one approach to understanding topics and public responses to dengue fever (Prabaswara & Saputra, 2020).

Considering previous research, this study compares the Support Vector Machine and Random Forest classification methods in analyzing dengue fever. Based on the background above, the following research questions emerge: What are the main topics and how do people respond to dengue fever on Twitter? To answer these questions, this study aims to identify frequently occurring topics in tweets related to dengue fever and analyze the public's responses based on the collected tweets.

2. Materials and Method

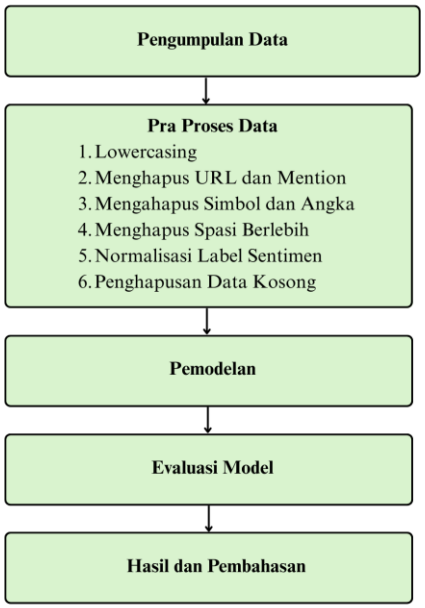

Figure 1. Research Method Flow.

Data Collection

This study used data collected through web crawling on Twitter using the keyword "Dengue Fever." The data collection process was carried out using the Python library snscrape, which automatically retrieves tweet data based on the keywords entered. Each collected tweet included information about the tweet text and posting time. After collection, the data was saved in .csv format and then manually labeled with sentiment by the researcher.

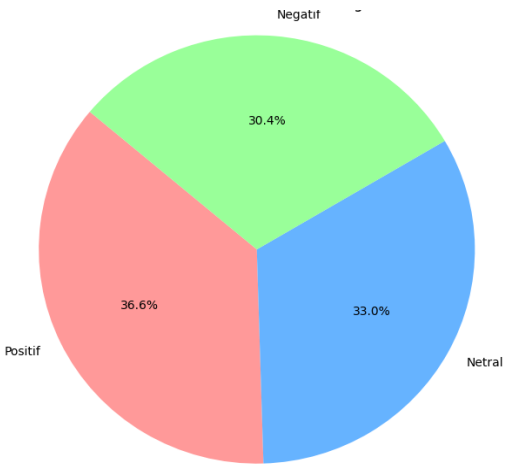

Figure 2. Data Sentiment Pie Chart.

Data Preprocessing

Data preprocessing is performed after the tweet data has been successfully collected to ensure the quality of the data used in the analysis. This data preprocessing stage aims to clean the text of irrelevant elements and prepare the data for processing by the classification model being designed (Riyandona et al., 2025).

a. Lowercasing
   Converting text to lowercase so that the model does not distinguish between the same word with different capitalizations (Aji Dewo Pangestu, 2025).
b. Removing URLs and Mentions
   Links and mentions are removed because they do not contribute to the meaning of the sentiment (Haris Kurnia Sandi Harahap et al., 2025).
c. Removing Symbols and Numbers
   Using regular expression methods to clean up special characters such as punctuation, emoticons, numbers, and other symbols (Putri Amira Sumitro, 2021).
d. Removing Excess Spaces
   Spacing normalization is performed to ensure the text is neat and easy for machines to read (Raif et al., 2024).
e. Sentiment Label Normalization
   Normalization to correct potential typos during manual labeling (M. Irfan Raif, 2024).
f. Removing Blank Data
   Rows that do not have text or sentiment labels are removed from the dataset to avoid errors during model training (Agung et al., 2023).

Modeling

The classification process was performed using two algorithms: Support Vector Machine (SVM) and Random Forest (RF). The dataset of 1,000 tweets was divided into 80% training data and 20% testing data using a stratified split to maintain a balanced label ratio. Both models were trained to classify the sentiment of tweets with the keyword "Dengue Fever," which were divided into three categories: Positive, Negative, and Neutral.

Model Evaluation

Evaluation was performed using accuracy, precision, recall, and f1-score metrics. The model performance comparison results were visualized as a bar graph with Positive, Negative, and Neutral sentiment categories. A macro average value was calculated to assess the model's average performance regardless of the number of data points per class. All evaluations were conducted using the Python library, and the results are displayed in tabular form to facilitate interpretation.

3. Results and Discussion

Support Vector Machine

Support Vector Machine (SVM) is a classification method using machine learning with a supervised learning approach. This method was first developed by Vladimir Vapnik and is used to predict classes based on patterns in data presentation (Aisah et al., 2023). Furthermore, SVM technically works by finding the best hyperplane that optimally separates two data classes. The goal of this separation is to maximize the margin or distance between data points from each class and the hyperplane. SVM has several advantages, including working effectively in high-dimensional spaces, being efficient in memory usage, and being resistant to overfitting, especially when the number of features exceeds the amount of data (Khoirul Abbi Rokhman, 2021). In the context of sentiment analysis, SVM is widely used because it can classify text data, such as tweets, with high accuracy. This is due to its ability to handle non-linear data using kernel tricks, making it particularly suitable for processing public opinion based on social media like Twitter.

The following are the results of Precision, recall, f1-score, and support obtained by the Support Vector Machine algorithm based on tests run through Google Colab.

```
=== SVM ===
Akurasi: 0.7707317073170732
              precision    recall  f1-score   support

     Negatif       0.78      0.69      0.73        65
      Netral       0.72      0.75      0.74        65
     Positif       0.81      0.85      0.83        75

    accuracy                           0.77       205
   macro avg       0.77      0.77      0.77       205
weighted avg       0.77      0.77      0.77       205
```

Figure 3. Support Vector Machine Model Evaluation Results.

The accuracy of the Support Vector Machine algorithm on the test data was 0.77 or 77.07%, with a precision of 0.77, a recall of 0.77, an f1-score of 0.77, and support for 205 test data points.

Random Forest

Random Forest is an ensemble learning-based classification method developed by Leo Breiman in 2001. This method consists of a set of decision trees randomly generated from a subset of disease data, then combining the results from each tree to determine the final prediction. Random Forest has advantages in handling large data sets, is resistant to overfitting, and performs well on data with many variables and noise. In the classification process, Random Forest uses a voting approach from all formed trees to determine the final class of the data (Suci Amaliah et al., 2022). In text-based sentiment analysis, such as tweets, Random Forest is able to handle various text features well and provides stable classification results. Furthermore, Random Forest can measure feature importance.

The following are the results of Precision, recall, f1-score, and support obtained by the Random Forest algorithm based on tests run through Google Colab.

```
=== Random Forest ===
Akurasi: 0.7365853658536585
              precision    recall  f1-score   support

     Negatif       0.77      0.63      0.69        65
      Netral       0.64      0.77      0.70        65
     Positif       0.81      0.80      0.81        75

    accuracy                           0.74       205
   macro avg       0.74      0.73      0.73       205
weighted avg       0.75      0.74      0.74       205
```

Figure 4. Random Forest Model Evaluation Results.

The Random Forest algorithm achieved an accuracy of 0.736 or 73.65%, with a precision of 0.74, a recall of 0.73, an f1-score of 0.73, and support for 205 test datasets.

Evaluation and Comparison

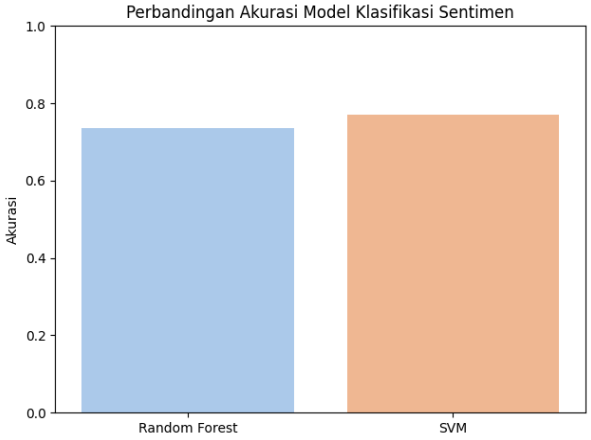

Figure 5. Comparison Graph of Accuracy of SVM and Random Forest Models.

Based on the evaluation results, the Support Vector Machine algorithm had a 3.42% superior accuracy compared to the Random Forest algorithm. These results indicate that the Support Vector Machine algorithm has a 3.42% superiority in sentiment classification for short Indonesian texts. In this study, sentiment analysis was conducted on the topic of dengue fever on social media platform X.

4. Conclusion

Based on the results of research conducted on tweet objects obtained from X, using the Support Vector Machine classification algorithm, and Random Forest used to classify sentiments on tweet objects with the keyword "dengue fever" it can be concluded that the SVM algorithm has superior performance compared to Random Forest, with an average of 77.07%, while Random Forest reached 73.65%. This shows that SVM is more effective and consistent in classifying public opinion from short text data such as tweets. The lack of this research as well as being part of the evaluation of this research is that a larger amount of data will help the model to recognize sentiments from tweet texts better, the application of deep learning is also able to make the model built even better.

References

Agung, A., Daniswara, A., Kadek, I., & Nuryana, D. (2023). Data preprocessing pola pada penilaian mahasiswa program profesi guru. *Journal of Informatics and Computer Science, 05*.

Aisah, I. S., Irawan, B., & Suprapti, T. (2023). Algoritma support vector machine (SVM) untuk analisis sentimen ulasan aplikasi Al Qur'an digital. *Jurnal Mahasiswa Teknik Informatika, 7*(6). https://doi.org/10.36040/jati.v7i6.8263

Aji Dewo Pangestu, L. S. H. (2025). Analisis sentimen terkait judi online di media sosial Instagram menggunakan Naïve Bayes. *https://rayyanjurnal.com/index.php/ijedr/article/view/4798*, 3(1). https://doi.org/10.57235/ijedr.v3i1.4798

Haris Kurnia Sandi Harahap, A., Haerani, E., Oktavia, L., & Kurnia, F. (2025). Klasifikasi sentimen masyarakat terhadap revisi undang-undang tentara nasional Indonesia menggunakan Naïve Bayes classifier. *Bulletin of Computer Science Research, 5*(4), 594-603. https://doi.org/10.47065/bulletincsr.v5i4.615

I Gede Willy Karya Mahardika, M. R. I. N. A. (2023). Hubungan pengetahuan ibu dengan perilaku pencegahan DBD pada anak usia sekolah di Desa Tegallengah. *https://doi.org/10.37294/jrkn.v7i1.473*, 7. https://doi.org/10.37294/jrkn.v7i1.473

Khoirul Abbi Rokhman, B. P. A. (2021). Perbandingan metode support vector machine dan decision tree untuk analisis sentimen review komentar pada aplikasi transportasi online. *https://doi.org/10.24076/joism.2021v3i1.341*, 2(2). https://doi.org/10.24076/joism.2021v3i1.341

M. Irfan Raif, N. N. H. T. M. (2024). Otomatisasi pendeteksi kata baku dan tidak baku pada data Twitter berbasis KBBI. *http://repositori.umrah.ac.id/id/eprint/7752*, 11(2). https://doi.org/10.25126/jtiik.20241127404

Prabaswara, I. R., & Saputra, R. (2020). Analisis data sosial media Twitter menggunakan Hadoop dan Spark. *IT Journal Research and Development, 4*(2). https://doi.org/10.25299/itjrd.2020.vol4(2).4099

Putri Amira Sumitro, R. D. I. M. W. S. (2021). Analisis sentimen terhadap vaksin Covid-19 di Indonesia pada Twitter menggunakan metode lexicon based. *https://doi.org/10.55377/j-icom.v2i2.4009*, 2(2). https://doi.org/10.33059/j-icom.v2i2.4009

Raif, M. I., Hidayati, N. N., & Matulatan, T. (2024). Otomatisasi pendeteksi kata baku dan tidak baku pada data Twitter berbasis KBBI. *Jurnal Teknologi Informasi dan Ilmu Komputer, 11*(2), 337-348. https://doi.org/10.25126/jtiik.20241127404

Ratna Dian K, I. R. A. S. (2022). Pemberantasan sarang nyamuk 3M PLUS dalam perspektif persepsi dan motivasi sebagai upaya pencegahan demam berdarah dengue. *https://doi.org/10.33657/jurkessia.v13i1.362*, XIII(1). https://doi.org/10.33657/jurkessia.v13i1.362

Restu Wulandari, A., Andarini, D., Idris, H., Anggraeni, R., Studi Ilmu Kesehatan Masyarakat, P., & Kesehatan Masyarakat Universitas Sriwijaya, F. (2024). Analisis faktor resiko perilaku masyarakat dalam pengendalian vektor pada kasus (DBD): Literatur review. *Jurnal Lentera Kesehatan Masyarakat, 3*(1), 35-44. https://jurnalkesmas.co.id/index.php/jlkm

Riyandona, S. A., Rahaningsih, N., Dana, R. D., & Mulyawan, -. (2025). Implementasi model analisis sentimen terhadap grup musik BTS menggunakan metode Naïve Bayes. *Jurnal Informatika dan Teknik Elektro Terapan, 13*(1). https://doi.org/10.23960/jitet.v13i1.5816

Suci Amaliah, Nusrang, M., & Aswi, A. (2022). Penerapan metode random forest untuk klasifikasi varian minuman kopi di kedai kopi Konijiwa Bantaeng. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research, 4*(3), 121-127. https://doi.org/10.35580/variansiunm31

WHO. (2024, May). Dengue - Global situation. *https://www.who.int/emergencies/disease-outbreak-news/item/2024-DON518?utm_source*.