Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 6 Iss. 4 October 2026, pp: 1960-1969
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Comparison of Machine Learning Algorithms for Multilingual Health Hoax Classification on X and YouTube

Endin Bella Sefia[1*], Eka Sahputra[2], Maie Istighosah[3]

[1,3]Department of Informatics, Telkom University, Purwokerto, Indonesia
[2]Department of Data Science, Telkom University, Purwokerto, Indonesia

E-Mail: [1]endinbellla@student.telkomuniversity.ac.id, [2]ekasahputra@telkomuniversity.ac.id, [3]maieistigh@telkomuniversity.ac.id

Received May 30th 2026; Revised Jun 09th 2026; Accepted Jul 07th 2026; Available Online Oct 30th 2026
Corresponding Author: Endin Bella Sefia
Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

Short health-related messages on social media can influence how users interpret diseases, vaccines, and treatment issues. Because this type of content often appears as brief posts or video titles, automatic classification is needed to help identify misleading claims. This research evaluates three machine learning algorithms, namely Naïve Bayes, Support Vector Machine, and Logistic Regression, for health-hoax classification on X and YouTube. The data were collected from January 2020 to January 2026 and consisted of Indonesian and English records. English texts were translated into Indonesian so that all records could be processed in one language pipeline. The processing stages included text cleaning, normalization, tokenization, stemming, stopword removal, IndoBERT-based pseudo-labeling, TF-IDF feature representation, Random Under Sampling, and parameter tuning using GridSearchCV. The final dataset contained 10,983 records, consisting of 7,157 non-hoax and 3,826 hoax records. The experimental results show that Support Vector Machine produced the highest weighted F1-score of 0.7467 in the 70:30 split scenario without balancing. Logistic Regression produced a close weighted F1-score of 0.7462 and obtained the highest AUC value of 0.8010. Random Under Sampling improved hoax recall, but it also reduced the overall weighted F1-score.

Keywords: Classification, Health Hoax, Machine Learning, Multilingual, Youtube

1. INTRODUCTION

Health information is increasingly consumed through digital platforms. Many users do not only rely on official medical sources, but also encounter health-related claims through short posts, video titles, comments, and shared links. X and YouTube have different content characteristics, but both can accelerate the spread of health information. X is commonly used for short reactions and public discussions, while YouTube may present health claims through titles and video narratives. This rapid circulation can support public awareness, but it can also increase the exposure of misleading claims before verification is carried out. Previous studies explain that social media influences the production, distribution, and reception of information in society [1]. In health communication, misinformation may affect public understanding, risk perception, trust in health authorities, and decisions related to prevention or treatment [2], [3], [4].

In Indonesia, digital hoaxes remain a serious communication issue. The Ministry of Communication and Informatics reported that 12,547 hoax issues had been handled by the end of 2023, including 2,357 health-related issues [5]. In the first quarter of 2023, 425 hoax issues were also identified across digital platforms [6]. These numbers show that health hoaxes continue to appear in online spaces and need a systematic detection approach. The problem is also found in different platform formats. X often contains brief opinions and responses to trending issues, while YouTube can expose users to health claims through video titles and related narratives. Previous research on COVID-19 vaccination videos found that several highly viewed YouTube videos contained information that was not aligned with factual evidence [7]. This indicates that misleading health content can appear in content with wide public reach.

Classifying health hoaxes from social media text is not straightforward. The text is usually short, informal, and inconsistent. Posts and video titles may contain slang, abbreviations, hashtags, hyperlinks, symbols, repeated characters, mixed-language expressions, and incomplete sentence structures. These characteristics make raw text unsuitable for direct use in machine learning models. Before classification, the text must be cleaned and converted into a more consistent format [8]. Natural Language Processing (NLP)

supports this preparation stage, while Term Frequency-Inverse Document Frequency (TF-IDF) can transform textual terms into numerical features that can be processed by classification algorithms [8], [9].

Previous research has applied several conventional machine learning methods for hoax and fake-news detection. Febriyanty et al. compared Naïve Bayes and Support Vector Machine (SVM) for hoax news classification and reported that Naïve Bayes achieved higher accuracy than Support Vector Machine [10]. Irena and Setiawan used Decision Tree C4.5 for hoax identification on Twitter, with unigram features producing the best result [11]. Ariatmanto and Rifai showed that feature extraction plays an important role in fake-news classification, and TF-IDF contributed to strong performance when combined with Random Forest [12]. These findings suggest that conventional algorithms can still be useful when the feature representation is suitable for the characteristics of the text.

Deep learning and transformer-based approaches have also been explored in related studies. Fauzy and Setiawan applied Convolutional Neural Network (CNN) methods for fake-news detection on social media and obtained high accuracy [13]. Sibaroni et al. compared Long Short-Term Memory (LSTM), CNN, and hybrid CNN-LSTM for misinformation detection in Indonesian social media [14]. Sinapoy et al. compared LSTM and IndoBERT for hoax identification on Twitter, and IndoBERT showed better performance [15]. Although these methods can produce strong results, they generally require more complex model architecture and higher computational resources. For practical classification scenarios, simpler algorithms such as Naïve Bayes, Support Vector Machine, and Logistic Regression remain relevant to evaluate.

Several international studies have also emphasized the relevance of machine learning models in fake-news and text classification tasks. Zhou and Zafarani reviewed fake-news detection methods and explained that textual content, writing style, propagation patterns, and source credibility are important aspects in automatic detection [16]. Ahmed et al. compared several text-classification techniques for fake-content detection and showed that feature extraction and classifier selection strongly affect model performance [17]. These studies indicate that conventional machine learning models remain relevant for text classification when supported by appropriate preprocessing, feature representation, and evaluation metrics.

Class imbalance is another factor that can influence classification results. When one class has more records than another, the model may learn more strongly from the majority class [18]. Random Under Sampling (RUS) can reduce the number of majority-class records in the training data so that the class distribution becomes more balanced [19]. However, balancing may also change model behavior. It can increase the detection of minority-class records, but it may also increase incorrect predictions for the majority class. Therefore, the effect of Random Under Sampling needs to be observed together with the performance of each algorithm.

This research focuses on multilingual health-hoax classification using data from X and YouTube. The dataset includes X posts and YouTube video titles in Indonesian and English from January 2020 to January 2026. English records are translated into Indonesian to keep the classification process consistent. Three algorithms are compared: Naïve Bayes, Support Vector Machine, and Logistic Regression. The classification pipeline includes preprocessing, IndoBERT-based pseudo-labeling, TF-IDF feature extraction with unigram and bigram settings, Random Under Sampling, and GridSearchCV-based parameter tuning.

The objective of this research is to compare the performance of Naïve Bayes, Support Vector Machine, and Logistic Regression in separating health-related content into hoax and non-hoax classes. The evaluation uses accuracy, precision, recall, F1-score, confusion matrix, and Area Under the Curve (AUC). This research also observes the effect of Random Under Sampling and feature extraction settings on model performance. The results are expected to provide a practical overview of conventional machine learning for multilingual health-hoax detection on social media.

The novelty of this research lies in the combination of multilingual health-hoax data, multiplatform social media sources, and comparative conventional machine learning evaluation in one consistent classification pipeline. Unlike previous studies that mainly focused on a single platform, one language, or one classification approach, this research uses Indonesian and English data from X and YouTube, translates English records into Indonesian, applies IndoBERT-based pseudo-labeling, and compares Naïve Bayes, Support Vector Machine, and Logistic Regression using TF-IDF features. The main contributions of this research are: first, providing a multilingual health-hoax classification pipeline for short social media texts; second, comparing three conventional machine learning algorithms under several data-splitting scenarios; third, analyzing the effect of Random Under Sampling on hoax-class detection; and fourth, showing the trade-off between overall weighted F1-score and hoax recall in imbalanced health-hoax classification.

2. MATERIALS AND METHOD

This research was designed as a supervised text classification experiment. The process began with collecting health-related data from X and YouTube, followed by data integration, translation, exploratory analysis, preprocessing, semi-automatic labeling, feature extraction, model training, parameter tuning, and evaluation. The research workflow is presented in Figure 1.

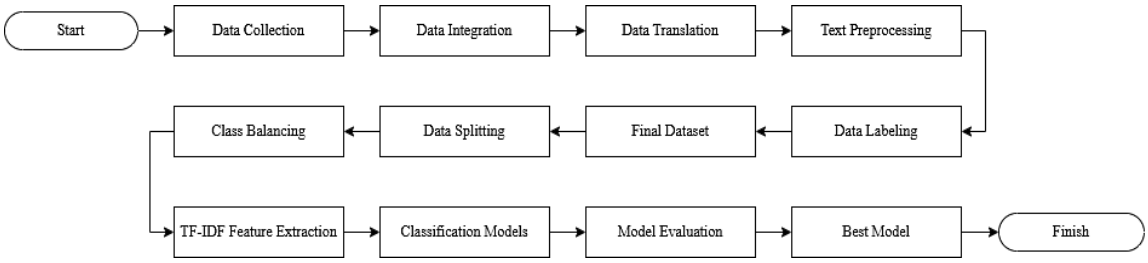


Figure 1. Research Methodology

Figure 1 illustrates the sequence of the research process. Data obtained from X and YouTube were merged into one dataset. Since the records contained Indonesian and English texts, the English records were translated into Indonesian before the classification process. The text was then prepared through preprocessing. IndoBERT was used to generate pseudo-labels, and invalid records were removed to obtain the final dataset. The experimental stage consisted of data splitting, class balancing, TF-IDF feature extraction, and model training using Naïve Bayes, Support Vector Machine, and Logistic Regression. After training, the models were evaluated and compared using weighted F1-score as the main selection metric.

2.1. Dataset

The dataset contains health-related text collected from X and YouTube. The records consist of X posts and YouTube video titles written in Indonesian and English. The collection period covered January 2020 to January 2026. At the initial stage, 12,626 records were obtained. After empty records and duplicate entries were removed, the merged dataset contained 11,019 records. English records were translated into Indonesian to make the language format consistent. After preprocessing and invalid-record filtering, 10,983 records were retained as the final dataset. The final data consisted of 7,157 non-hoax records and 3,826 hoax records. The dataset summary is shown in Table 1.

Table 1. Dataset Summary

Stage	Number of Records
Initial collected data	12,626
Data after removing empty records	12,621
Removed duplicate records	1,602
Merged dataset	11,019
Final dataset for classification	10,983
Non-hoax class	7,157
Hoax class	3,826

2.2. Data Preparation and Text Preprocessing

Data preparation was conducted to make the collected records suitable for classification. This stage included merging data from different sources, removing empty entries, checking duplicate records, standardizing columns, and translating English texts into Indonesian. These steps were needed because the data came from different platforms and language sources. By using Indonesian as the processing language, all records could be handled through one classification pipeline.

Preprocessing was then applied to improve text consistency and reduce noise [8]. The steps included cleaning, case folding, word normalization, tokenization, stemming, stopword removal, and clean-text formation. Cleaning removed URLs, symbols, punctuation, and irrelevant characters. Case folding changed uppercase letters into lowercase. Word normalization converted informal expressions into standard forms. Tokenization separated text into word units. Stemming reduced words to their base forms, while stopword removal deleted common words with low classification value. The output of this stage was used as input for feature extraction.

2.3. Data Labeling

The labeling process used IndoBERT, which has been applied in previous Indonesian hoax identification research [15]. Each record was categorized into hoax or non-hoax. Label 1 was assigned to hoax records, while label 0 was assigned to non-hoax records. Because the labels were generated automatically by a model and not manually determined by human annotators, they are treated as pseudo-labels.

A confidence score analysis was conducted to review the labeling output. This analysis was used to observe the distribution of IndoBERT scores, identify high-confidence records, and inspect records that may fall into ambiguous areas. After this stage, the labeled data were used as the target variable for model training and evaluation.

2.4. Feature Extraction

Feature extraction was performed using Term Frequency-Inverse Document Frequency. This method gives a numerical value to each term by considering how often the term appears in a document and how widely it appears across the corpus [9]. A term receives a higher weight when it is frequent in a specific document but less common in the entire document collection. The TF-IDF calculation is shown in Equation (1).

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \tag{1}$$

The inverse document frequency value is calculated using Equation (2).

$$IDF(t) = log\left(\frac{N+1}{(DF(t)+1)}\right) + 1 \tag{2}$$

In these equations, TF-IDF(*t*,*d*) denotes the weight of term *t* in document *d*. TF(*t*,*d*) denotes the occurrence frequency of term t in document d. IDF(*t*) denotes the inverse document frequency of term *t*. *N* refers to the total number of documents, while DF(*t*) refers to the number of documents containing term t. This research uses unigram and bigram as feature scenarios.

2.5. Experimental Scenarios

The final dataset was divided using stratified splitting to keep the class proportion consistent in each scenario. Four split ratios were tested: 60:40, 70:30, 80:20, and 90:10. Two class distribution conditions were compared. The first condition used the original class distribution, while the second condition used Random Under Sampling. RUS reduced majority-class records in the training set to create a more balanced class composition [19]. The experimental scenarios are shown in Table 2.

Table 2. Experimental Scenarios

Component	Scenario
Data splitting	60:40, 70:30, 80:20, 90:10
Class distribution	Without balancing and Random Under Sampling
Feature extraction	TF-IDF with unigram and bigram
Classification models	Naïve Bayes, Support Vector Machine, Logistic Regression
Optimization method	GridSearchCV with 5-fold cross-validation
Main selection metric	Weighted F1-score

2.6. Classification Models

Three algorithms were used in this research: Naïve Bayes, Support Vector Machine, and Logistic Regression. These algorithms were selected because they are widely used in text classification and can be combined with TF-IDF features. Naïve Bayes and Support Vector Machine have been applied in previous hoax classification research [10]. Logistic Regression was included because it is commonly used to model the relationship between input features and class labels [20].

Naïve Bayes classifies records using a probability calculation based on Bayes' theorem. This algorithm is often used in text classification because it can process many word-based features. Support Vector Machine separates classes by forming a decision boundary in the feature space. Logistic Regression estimates the probability that a record belongs to a certain class and assigns the record to the class with the highest probability.

All algorithms were trained using TF-IDF features. Hyperparameter tuning was conducted using GridSearchCV with 5-fold cross-validation. This procedure tested several parameter combinations and selected the configuration that produced the best score [21]. Weighted F1-score was used as the optimization metric because the dataset contained unequal numbers of hoax and non-hoax records. The tested parameters are summarized in Table 3.

Table 3. Classification Models and Tested Parameters

Model	Tested Parameter
Naïve Bayes	alpha
Support Vector Machine	C
Logistic Regression	C
TF-IDF	ngram_range, min_df, max_df

2.7. Evaluation Metrics

The evaluation stage used accuracy, precision, recall, F1-score, confusion matrix, and Area Under the Curve. These metrics were applied to measure how closely the model predictions matched the actual labels [22]. The confusion matrix used in this research is shown in Table 4.

Table 4. Confusion Matrix

Actual Class	Predicted Hoax	Predicted Non-Hoax
Hoax	True Positive (TP)	False Negative (FN)
Non-Hoax	False Positive (FP)	True Negative (TN)

In Table 4, True Positive refers to hoax records correctly classified as hoax. True Negative refers to non-hoax records correctly classified as non-hoax. False Positive occurs when non-hoax records are predicted as hoax. False Negative occurs when hoax records are predicted as non-hoax.

Accuracy shows the proportion of correct predictions from all testing records. Precision measures the correctness of records predicted as hoax. Recall shows how many actual hoax records are detected by the model. F1-score combines precision and recall so that both aspects can be considered in one value. The formulas are presented in Equations (3) to (6).

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{3}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{4}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{5}$$

$$F1 - score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{6}$$

TP denotes True Positive, TN denotes True Negative, FP denotes False Positive, and FN denotes False Negative. AUC was used to observe the model's ability to separate hoax and non-hoax classes through the ROC curve [23]. The best model was determined using weighted F1-score because this metric considers class distribution and is more suitable for imbalanced data than accuracy alone.

3. RESULTS AND DISCUSSION

This section reports the classification results of Naïve Bayes, Support Vector Machine, and Logistic Regression. The experiment used four data split ratios and two class distribution settings, namely the original distribution and Random Under Sampling. Weighted F1-score was used as the main basis for comparing model performance. Accuracy, precision, recall, confusion matrix, and AUC were used to support the analysis.

3.1. Overall Scenario Results

The first comparison was made by observing the weighted F1-score values from all experimental scenarios. The complete results are presented in Table 5.

Table 5. Weighted F1-Score Results for All Scenarios

Split	Balancing	Naïve Bayes	SVM	Logistic Regression
60:40	Without balancing	0.7167	0.7379	0.7393
	RUS	0.7217	0.7355	0.7322
70:30	Without balancing	0.7169	0.7467	0.7462
	RUS	0.7203	0.7244	0.7256
80:20	Without balancing	0.7088	0.7353	0.7373
	RUS	0.7067	0.7296	0.7369
90:10	Without balancing	0.7242	0.7349	0.7310
	RUS	0.7111	0.7122	0.7096

Based on Table 5, Support Vector Machine produced the strongest weighted F1-score. The best value was 0.7467, obtained from the 70:30 split scenario without balancing. Logistic Regression achieved a very close value in the same scenario, with a weighted F1-score of 0.7462. Naïve Bayes reached its highest weighted

F1-score of 0.7242 in the 90:10 split scenario without balancing. These results indicate that SVM and Logistic Regression were more effective than Naïve Bayes for the TF-IDF representation used in this research. The comparison of weighted F1-score values across all scenarios is displayed in Figure 2.

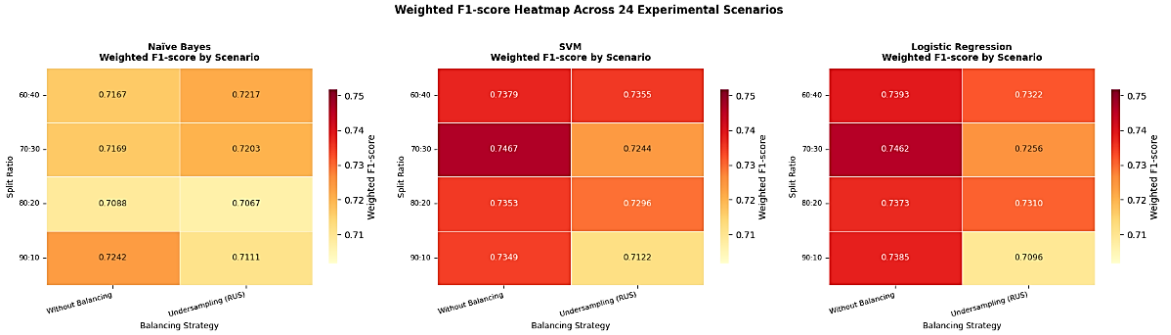


Figure 2. Weighted F1-score Comparison of All Scenarios

Figure 2 shows that the scenarios using the original class distribution generally produced better weighted F1-score values than the scenarios using Random Under Sampling. SVM and Logistic Regression showed similar performance patterns, especially in the 70:30 split without balancing. Naïve Bayes tended to produce lower values across most scenarios, indicating that it was less competitive in this experiment.

The results in Table 5 and Figure 2 indicate that model performance was influenced by both the split ratio and the balancing strategy. The 70:30 split without balancing provided the most stable condition for SVM and Logistic Regression because the training data were sufficient to learn discriminative TF-IDF patterns, while the testing data still represented the original class distribution. The lower performance of Naïve Bayes across most scenarios suggests that the independence assumption of this algorithm was less suitable for short and informal social media texts, where the meaning of a health claim may depend on word combinations and contextual patterns.

3.2. Best Scenario of Each Model

The best scenario for each algorithm was selected based on the highest weighted F1-score produced by that algorithm. The comparison is shown in Table 6.

Table 6. Best Scenario of Each Classification Model

Model	Split	Balancing	Accuracy	Precision	Recall	F1-score	Hoax Recall	Hoax F1-score
Naïve Bayes	90:10	Without Balancing	0.7343	0.7254	0.7343	0.7242	0.4987	0.5668
SVM	70:30	Without Balancing	0.7520	0.7459	0.7520	0.7467	0.5688	0.6152
Logistic Regression	70:30	Without Balancing	0.7530	0.7463	0.7530	0.7462	0.5531	0.6094

Table 6 shows that SVM obtained the highest weighted F1-score, with a value of 0.7467. Logistic Regression produced a slightly higher accuracy value of 0.7530, while SVM obtained 0.7520. However, accuracy was not used as the main criterion because the dataset was imbalanced. In this condition, a high accuracy value can still occur even when the minority class is not detected optimally.

For the hoax class, SVM achieved a recall of 0.5688 and a hoax F1-score of 0.6152. Logistic Regression obtained a hoax recall of 0.5531 and a hoax F1-score of 0.6094. These values show that SVM was slightly better at identifying hoax records in the best scenario without balancing. Naïve Bayes produced the lowest performance, with a weighted F1-score of 0.7242 and a hoax recall of 0.4987

3.3. Effect of Random Under Sampling

The influence of Random Under Sampling was analyzed by comparing the best scenario without balancing and the best scenario with RUS. The comparison is presented in Table 7.

Table 7 shows that RUS increased hoax recall in all three models. In Naïve Bayes, hoax recall increased from 0.4987 to 0.7231. In SVM, the value increased from 0.5688 to 0.7185. In Logistic Regression, the value increased from 0.5531 to 0.7211. This means that the models became more sensitive to hoax records after the training data were balanced.

Table 7. Comparison of Without Balancing and RUS Scenarios

Model	Balancing	Split	Accuracy	Precision	Recall	F1-score	Hoax Recall	Hoax F1-score
Naïve Bayes	Without Balancing	90:10	0,7343	0,7254	0,7343	0,7242	0,4987	0,5668
	RUS	60:40	0,7160	0,7392	0,7160	0,7217	0,7231	0,6395
SVM	Without Balancing	70:30	0,7520	0,7459	0,7520	0,7467	0,5688	0,6152
	RUS	60:40	0,7308	0,7480	0,7308	0,7355	0,7185	0,6503
Logistic Regression	Without Balancing	70:30	0,7530	0,7463	0,7530	0,7462	0,5531	0,6094
	RUS	60:40	0,7271	0,7461	0,7271	0,7322	0,7211	0,6481

However, the improvement in hoax recall was followed by a decrease in weighted F1-score. In SVM, the weighted F1-score decreased from 0.7467 to 0.7355. In Logistic Regression, the value decreased from 0.7462 to 0.7322. This result indicates a trade-off between detecting more hoax records and maintaining stable overall performance. Therefore, the scenario without balancing is more appropriate when the main target is overall stability. RUS can still be considered when the priority is to reduce missed hoax records, although the risk of false-positive predictions increases.

The improvement in hoax recall after applying RUS shows that class distribution strongly affected the model's sensitivity toward the minority class. By reducing the number of non-hoax records in the training set, the models received a more balanced representation of hoax and non-hoax classes. However, this condition also reduced some majority-class information, which may explain the decrease in weighted F1-score. Therefore, RUS did not simply improve all performance metrics, but shifted the model behavior toward detecting more hoax records at the cost of increasing false positive predictions.

3.4. Confusion Matrix and AUC Analysis

The confusion matrix was used to observe how predictions were distributed across correct and incorrect categories. The summary is shown in Table 8.

Table 8. Confusion Matrix Summary

Model	Scenario	TN	FP	FN	TP
Naïve Bayes	Without Balancing	616	100	192	191
	RUS	2,039	824	424	1,107
SVM	Without Balancing	1,825	322	495	653
	RUS	2,111	752	431	1,100
Logistic Regression	Without Balancing	1,846	301	513	635
	RUS	2,091	771	427	1,104

Table 8 shows that the scenarios without balancing generated fewer false positive predictions. This means that the models were less likely to classify non-hoax records as hoax. After RUS was applied, true positive predictions increased in SVM and Logistic Regression. In SVM, the true positive value increased to 1,100, while the false negative value decreased to 431. In Logistic Regression, the true positive value increased to 1,104, while the false negative value decreased to 427.

These results show that RUS helped the models detect more hoax records. However, this improvement also increased false positive predictions. In the scenario without balancing, SVM produced fewer false negatives than Logistic Regression, with 495 false negatives compared with 513. This supports the selection of SVM as the best model because it missed fewer hoax records while still maintaining the highest weighted F1-score. The AUC values of the best scenario for each model are presented in Table 9.

Table 9. AUC Values of the Best Scenario

Model	Split	Balancing	AUC
Naïve Bayes	90:10	Without balancing	0,7814
SVM	70:30	Without balancing	0,7981
Logistic Regression	70:30	Without balancing	0,8010

The ROC curve of the best classification models is presented in Figure 3.

Based on Table 9 and Figure 3, Logistic Regression produced the highest AUC value of 0.8010. SVM followed with an AUC value of 0.7981, while Naïve Bayes obtained 0.7814. The difference between Logistic Regression and SVM was only 0.0029, indicating that both models had nearly similar ability to separate hoax and non-hoax records.

Although Logistic Regression obtained the highest AUC value, SVM was considered more appropriate as the best model because the main evaluation metric in this research was weighted F1-score. In addition, SVM produced fewer false negative predictions than Logistic Regression in the best scenario. This is important in

health-hoax classification because false negative predictions may allow misleading health information to be treated as non-hoax content.

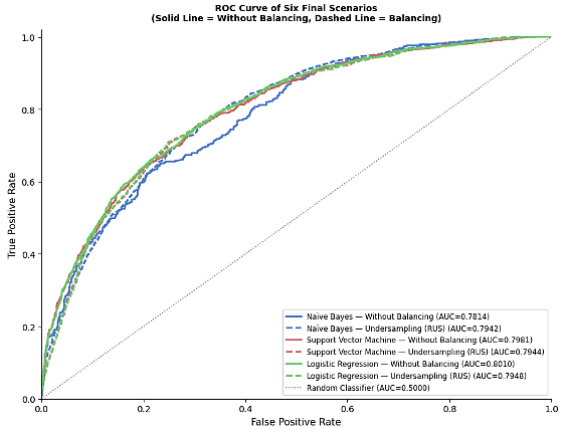


Figure 3. ROC Curve of the Best Classification Models

3.5. Discussion

The experimental results show that Support Vector Machine and Logistic Regression produced better performance than Naïve Bayes in multilingual health-hoax classification. This result is closely related to the characteristics of TF-IDF representation. TF-IDF produces sparse and high-dimensional features, and both SVM and Logistic Regression are generally suitable for this type of feature space. SVM works by finding an optimal separating boundary between classes, while Logistic Regression estimates class probability based on weighted feature combinations. These characteristics allow both models to capture more discriminative patterns from short social media texts. In contrast, Naïve Bayes assumes conditional independence among features. This assumption is simple and efficient, but it may be less suitable for informal social media text because the meaning of a claim is often influenced by word combinations, context, and phrase patterns.

The finding is consistent with previous machine learning studies in text classification, which show that classifier selection and feature extraction significantly influence classification performance. Previous fake-news and hoax-detection studies have reported that conventional algorithms such as Naïve Bayes, SVM, and Logistic Regression can be effective when combined with suitable textual features. However, the performance can vary depending on the dataset, topic, language, platform, and evaluation setting. Compared with previous studies that mainly used one platform or one language, this research expands the classification setting by using multilingual health-related data from X and YouTube. Therefore, the result contributes to the existing body of knowledge by showing that conventional machine learning models are still applicable in a multilingual and multiplatform health-hoax classification context.

The best overall result was obtained by SVM in the 70:30 split scenario without balancing, with a weighted F1-score of 0.7467. Logistic Regression achieved a very close weighted F1-score of 0.7462 and produced the highest AUC value of 0.8010. This small difference indicates that both models had almost similar ability to distinguish hoax and non-hoax records. However, SVM was selected as the best model because weighted F1-score was used as the main selection metric and because SVM produced fewer false negative predictions than Logistic Regression in the best scenario. In health-hoax classification, false negative errors are important because hoax records that are predicted as non-hoax may allow misleading health information to remain undetected.

The comparison between the original class distribution and Random Under Sampling shows that balancing strategy changes the behavior of the models. Without balancing, the models generally produced higher weighted F1-score values. This means that the models were more stable in maintaining overall performance across the dataset distribution. After RUS was applied, hoax recall increased in all three models. Naïve Bayes increased from 0.4987 to 0.7231, SVM increased from 0.5688 to 0.7185, and Logistic Regression increased from 0.5531 to 0.7211. These results indicate that RUS made the models more sensitive to the hoax class by reducing the dominance of the non-hoax class during training.

However, the improvement in hoax recall was followed by a decrease in weighted F1-score and an increase in false positive predictions. This trade-off shows that RUS can help detect more hoax records, but it also increases the possibility of classifying non-hoax records as hoax. In practical use, this trade-off should be considered carefully. If the purpose of the system is to maintain stable overall classification performance, SVM without balancing is more appropriate. If the purpose is to detect as many hoax records as possible, SVM with RUS can be considered, although the increase in false positive predictions must be managed.

The results also show that the proposed pipeline has practical implications for early health-hoax identification on social media. The use of preprocessing, translation, IndoBERT-based pseudo-labeling, TF-IDF, and conventional machine learning provides a relatively lightweight classification approach compared with more complex deep learning models. This can be useful when computational resources are limited or when a simple baseline model is needed before developing more advanced systems. Nevertheless, this research has several limitations. The labels were generated using IndoBERT and treated as pseudo-labels, so manual annotation by experts or multiple annotators is still needed to improve label reliability. In addition, YouTube data were limited to video titles, while video content, comments, and metadata were not analyzed. Future research can address these limitations by expanding data sources, using verified fact-checking references, involving human annotators, and comparing more advanced models such as XGBoost, LSTM, IndoBERT fine-tuning, or other transformer-based architectures.

4. CONCLUSION

This research compared Naïve Bayes, Support Vector Machine, and Logistic Regression for classifying health-related social media text into hoax and non-hoax classes. The data were collected from X and YouTube and consisted of Indonesian and English records. English records were translated into Indonesian, and the data were processed through preprocessing, IndoBERT-based pseudo-labeling, TF-IDF feature extraction, class balancing, and hyperparameter tuning.

The best overall result was produced by Support Vector Machine in the 70:30 split scenario without balancing. This model achieved a weighted F1-score of 0.7467. Logistic Regression produced a very close weighted F1-score of 0.7462 and obtained the highest AUC value of 0.8010. Nevertheless, SVM was selected as the best model because it achieved the highest weighted F1-score and produced fewer false negative predictions than Logistic Regression in the best scenario.

Random Under Sampling increased hoax recall in all three algorithms. In Naïve Bayes, hoax recall increased from 0.4987 to 0.7231. In SVM, the value increased from 0.5688 to 0.7185. In Logistic Regression, the value increased from 0.5531 to 0.7211. However, the increase in hoax recall was followed by lower weighted F1-score values and more false positive predictions. Therefore, the scenario without balancing is more suitable when stable overall performance is the main objective. RUS can be used when detecting more hoax records is considered more important, although it may increase false positive predictions.

Future research can expand the dataset by adding other social media platforms and verified fact-checking sources. Manual annotation involving experts or multiple annotators is also recommended to improve label reliability. Further studies may compare other balancing methods and more advanced models, such as Random Forest, XGBoost, LSTM, IndoBERT, or other transformer-based models.

REFERENCES

[1] E. Zhuravskaya, M. Petrova, and R. Enikolopov, “Political Effects of the Internet and Social Media,” *Annu. Rev. Econom.*, vol. 12, no. 1, pp. 415–438, Aug. 2020, doi: 10.1146/annurev-economics-081919-050239.

[2] G. K. Shahi, A. Dirkson, and T. A. Majchrzak, “An exploratory study of COVID-19 misinformation on Twitter,” *Online Soc. Netw. Media*, vol. 22, no. May 2020, p. 100104, Mar. 2021, doi: 10.1016/j.osnem.2020.100104.

[3] W. H. Organization, “Infodemic,” World Health Organization. Accessed: May 05, 2026. [Online]. Available: https://www.who.int/health-topics/infodemic

[4] W. H. O. R. O. for Europe, “Infodemics and misinformation negatively affect people’s health behaviours, new WHO review finds,” World Health Organization. Accessed: May 05, 2026. [Online]. Available: https://www.who.int/europe/news/item/01-09-2022-infodemics-and-misinformation-negatively-affect-people-s-health-behaviours--new-who-review-finds

[5] Kementerian Komunikasi dan Informatika Republik Indonesia, “Hingga akhir tahun 2023, Kominfo tangani 12.547 isu hoaks.” Accessed: May 09, 2025. [Online]. Available: https://www.komdigi.go.id/berita/pengumuman/detail/siaran-pers-no-02-hm-kominfo-01-2024-tentang-hingga-akhir-tahun-2023-kominfo-tangani-12-547-isu-hoaks

[6] K. K. dan I. R. Indonesia, “Triwulan pertama 2023: Kominfo identifikasi 425 isu hoaks.” Accessed: May 09, 2025. [Online]. Available: https://www.komdigi.go.id/berita/siaran-pers/detail/siaran-pers-no-50-hm-kominfo-04-2023-tentang-triwulan-pertama-2023-kominfo-identifikasi-425-isu-hoaks

[7] H. Oi-, Y. Li, E. Pastukhova, O. Brandts-, M. G. Tan, and M. G. Kirchhof, “YouTube as a source of misinformation on COVID- 19 vaccination : a systematic analysis,” no. November 2021, pp. 1–6, 2022, doi: 10.1136/bmjgh-2021-008334.

[8] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, “Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations,” *Organ. Res. Methods*, vol. 25, no. 1, pp. 114–146, Jan. 2022, doi: 10.1177/1094428120971683.

[9] Z. Zhan, "Comparative Analysis of TF-IDF and Word2Vec in Sentiment Analysis: A Case of Food Reviews," *ITM Web of Conferences*, vol. 70, p. 02013, Jan. 2025, doi: 10.1051/itmconf/20257002013.
[10] N. E. Febriyanty, M. A. Hariyadi, and C. Crysdian, "Hoax Detection News Using Naïve Bayes and Support Vector Machine Algorithm," *International Journal of Advances in Data and Information Systems*, vol. 4, no. 2, pp. 191–200, Oct. 2023, doi: 10.25008/ijadis.v4i2.1306.
[11] B. Irena and Erwin Budi Setiawan, "Fake News (Hoax) Identification on Social Media Twitter using Decision Tree C4.5 Method," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 4, pp. 711–716, Aug. 2020, doi: 10.29207/resti.v4i4.2125.
[12] D. Ariatmanto and A. M. Rifai, "The Impact of Feature Extraction in Random Forest Classifier for Fake News Detection," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 6, pp. 730–736, Dec. 2024, doi: 10.29207/resti.v8i6.6017.
[13] A. R. I. Fauzy and Erwin Budi Setiawan, "Detecting Fake News on Social Media Combined with the CNN Methods," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 2, pp. 271–277, Mar. 2023, doi: 10.29207/resti.v7i2.4889.
[14] Y. Sibaroni, S. Mahadzir, S. S. Prasetiyowati, and A. F. Ihsan, "Combating Misinformation: Leveraging Deep Learning for Hoax Detection in Indonesian Political Social Media," *JURNAL INFOTEL*, vol. 16, no. 2, pp. 413–426, May 2024, doi: 10.20895/infotel.v16i2.1139.
[15] Muhammad Ikram Kaer Sinapoy, Yuliant Sibaroni, and Sri Suryani Prasetyowati, "Comparison of LSTM and IndoBERT Method in Identifying Hoax on Twitter," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 3, pp. 657–662, 2023, doi: 10.29207/resti.v7i3.4830.
[16] X. Zhou and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," *ACM Comput. Surv.*, vol. 53, no. 5, 2021, doi: 10.1145/3395046.
[17] H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," *Security and Privacy*, vol. 1, no. 1, pp. 1–15, 2018, doi: 10.1002/spy2.9.
[18] Y. Yang, H. A. Khorshidi, and U. Aickelin, "A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems," *Front. Digit. Health*, vol. 6, no. July, 2024, doi: 10.3389/fdgth.2024.1430245.
[19] J. Hoyos-Osorio, A. Alvarez-Meza, G. Daza-Santacoloma, A. Orozco-Gutierrez, and G. Castellanos-Dominguez, "Relevant information undersampling to support imbalanced data classification," *Neurocomputing*, vol. 436, pp. 136–146, 2021, doi: 10.1016/j.neucom.2021.01.033.
[20] S. U. Hassan, J. Ahamed, and K. Ahmad, "Analytics of machine learning-based algorithms for text classification," *Sustainable Operations and Computers*, vol. 3, no. March, pp. 238–248, 2022, doi: 10.1016/j.susoc.2022.03.001.
[21] S. F. M. Radzi, M. K. A. Karim, M. I. Saripan, M. A. A. Rahman, I. N. C. Isa, and M. J. Ibahim, "Hyperparameter tuning and pipeline optimization via grid search method and tree-based autoML in breast cancer prediction," *J. Pers. Med.*, vol. 11, no. 10, 2021, doi: 10.3390/jpm11100978.
[22] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, 2018, doi: 10.1016/j.aci.2018.08.003.
[23] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.