Implementation of Random Forest Algorithm with Random Oversampling for Sentiment Analysis of X Users Toward the *Sekolah Rakyat* Program

Deni Agus Hendrawan[1], Ali Ibrahim[2*], Yadi Utama[3], Endang Lestari Ruskan[4], Fathoni[5]

[1,3,5]Information System, Faculty of Computer, Sriwijaya University, Palembang, South Sumatera, Indonesia
[2]Master of Computer Science, Faculty of Computer, Sriwijaya University, Palembang, South Sumatera, Indonesia
[4]Computer System, Faculty of Computer, Sriwijaya University, Palembang, South Sumatera, Indonesia
E-mail: [1]deniadeni256@gmail.com, [2*]aliibrahim@unsri.ac.id, [3]yadiutama@unsri.ac.id, [4]endanglestari@unsri.ac.id, [5]fathoni@unsri.ac.id

(Received: 21 Jan 2026, revised: 22 Feb 2026, 18 Mar 2026, accepted: 26 Mar 2026)

Abstract

Social media site X has emerged as a significant platform for voicing public views on government initiatives, such as the *Sekolah Rakyat* Program. Nevertheless, utilizing social media information for sentiment analysis often faces challenges due to class imbalance, which may result in skewed predictions from models. This research seeks to examine public sentiment and assess how well the Random Forest algorithm performs when paired with Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction and Random Oversampling (ROS) methods to mitigate class imbalance. A dataset comprising 8,623 tweets was gathered and split into training and testing sets using an 80:20 ratio. The results of the experiments indicate that the suggested method demonstrates robust and realistic classification performance, achieving an accuracy of 80.99%, along with a weighted average score in precision, recall, and F1-score of 0.81. Additionally, the sentiment analysis indicates that the majority of public opinions are largely positive, with roughly 69.4% of the testing data reflecting a favorable outlook toward free education access and school improvement efforts. These findings suggest that the proposed model provides reliable performance in capturing public sentiment patterns.

Keywords: Random Forest, Random Oversampling, *Sekolah Rakyat* Program, Sentiment Analysis, Social Media X

I. INTRODUCTION

Social media is now a key source of information, helping us figure out what people think about government actions. Platform X lets people share what they think quickly, widely, and without holding back, allowing researchers to examine feelings as a way to measure how well government policies are working [1], [2]. Examining emotions using social media is seen as useful because it can quickly show what people are feeling and changing views.

Within the field of public policy, leveraging data obtained from social media platforms has become more and more common for assessing how the public feels about government initiatives. The views that people share on social media can give us initial clues about how well a policy is doing and what problems might arise as it's being put into action [3]. As a result, understanding the emotions expressed in these opinions is vital for making choices that are based on solid data.

The *Sekolah Rakyat* Program, an initiative focused on widening educational opportunities for less fortunate communities, has garnered considerable notice from the general population. The reactions from the public regarding this program have varied greatly, spanning from approval to disapproval, with these opinions being broadly circulated on the X platform. Consequently, examining the feelings of the public is of utmost importance as a foundation for both judging and refining educational strategies [4], [5].

However, social networking site details usually lack a well-defined structure and are filled with considerable amounts of irrelevant content, like casual speech, shortened terms, and diverse word choices. These traits create difficulties for assessing feelings expressed in text, since they might impact how good the data features are and how well the text can be sorted if these issues are not correctly managed [6].

Sentiment analysis performed on social media often struggles with class imbalances alongside challenges posed by unstructured data. In most cases, a particular sentiment

category is far more prevalent than others in the dataset, which leads classification models to favor the more frequent category, thereby overlooking less represented ones [7], [8]. This imbalance can compromise the accuracy and trustworthiness of sentiment analysis results.

Numerous machine learning techniques have been used to tackle issues in sentiment analysis, with the Random Forest algorithm being one such method. The algorithm stands out because of its notable ability to generalize well, its resilience to overfitting, and its aptitude for processing complex datasets such as those containing text [9], [10]. Research indicates that Random Forest delivers more consistent and dependable results when compared against other classification techniques.

However, the effectiveness of Random Forest might diminish when used on datasets with uneven class distributions. Consequently, strategies for balancing data are necessary to improve the model's skill in identifying less frequent classes. Random Oversampling is chosen as it boosts the count of samples in the minority class, maintaining the integrity of the initial data traits and offering simple execution [11], [12].

Besides classification algorithms, methods for extracting features are very important in sentiment analysis that uses text. A common technique is Term Frequency–Inverse Document Frequency (TF-IDF) since it does a good job showing how important a word is in a single document compared to all the documents, and it works well with different machine learning algorithms [13].

This research, considering the aforementioned context, combines TF-IDF for identifying important features, Random Oversampling to handle data distribution, and the Random Forest algorithm to understand what people think about the *Sekolah Rakyat* Program on platform X. This study builds upon earlier work by specifically tackling the problem of uneven data distribution when looking at education policy in Indonesia, offering a fairer and more precise view of what the public thinks [14].

The aim of this research is to examine how the public feels about the *Sekolah Rakyat* Program by looking at information gathered from social media platform X, and to assess how well the Random Forest method, along with Random Oversampling and TF-IDF, works when dealing with unbalanced data categories.

The scope of this research focuses exclusively on Indonesian tweets concerning the *Sekolah Rakyat* Program, where sentiment categorization is split into positive and negative classes. The framework is limited to utilizing the Random Forest algorithm alongside Random Oversampling, and the assessment of performance is done through measures such as accuracy, precision, recall, and the F1-score.

This study aims to offer academic input for advancing machine learning methodologies in sentiment analysis, specifically for datasets with uneven class distribution, and to offer concrete advantages to decision-makers as an evaluative resource for the *Sekolah Rakyat* Program, derived from public sentiments shared on social platforms.

II. METHOD

For this research, the Random Forest method, together with Random Oversampling (ROS) and TF-IDF weighting, was used to precisely analyze user sentiments about the *Sekolah Rakyat* Program on platform X. The complete research process can be seen in Figure 1.

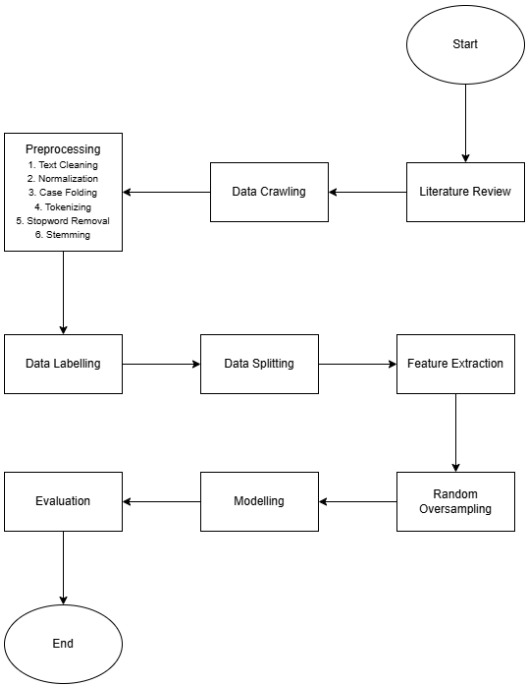


Figure 1. Research Flow

A. Data Crawling

The dataset used in this study was collected from the social media platform X (formerly Twitter), which has been widely utilized in sentiment analysis research due to its openness and real-time public discourse characteristics [15]. The data acquisition process was conducted using web scraping techniques supported by appropriate programming tools [16]. Tweets were retrieved using the keyword "*Sekolah Rakyat*," covering the period from January 2025 to December 2025. A total of 8,623 tweets were successfully collected and stored in CSV format for further preprocessing and analysis.

B. Preprocessing

Raw social media data are typically unstructured and contain noise, including informal expressions, symbols, and irrelevant textual elements. Therefore, several preprocessing steps were performed to improve data quality prior to analysis [17]. The preprocessing stages include:

a. Cleaning:
   Removing irrelevant elements from the text such as emojis, symbols, numbers, usernames, URLs, and hashtags using regular expression (Regex) methods.
b. Case Folding:
   Converting all characters in the text to lowercase to standardize the data format.
c. Normalization:

Transforming non-standard or slang words into their standard forms according to Indonesian language rules using a normalization dictionary.

d. Tokenizing:
Splitting sentences into individual word units (tokens).

e. Stopword Removal:
Eliminating commonly occurring words that do not carry significant sentiment meaning (such as “dan”, “yang”, and “di”) using the Sastrawi library.

f. Stemming:
Reducing inflected words to their root forms to unify different word variations that convey the same meaning.

C. Data Labelling

Following the data cleanup, a labeling process was executed to categorize every tweet into one of two sentiment groups: either positive or negative. Prior to the classification modeling phase, the ground-truth sentiment labels for the dataset were generated using a lexicon-based analytical approach. This process involved the utilization of two specialized Indonesian sentiment dictionaries containing positive and negative terms along with their respective polarity weights. For every preprocessed tweet, a cumulative polarity score was computed by aggregating the weights of its constituent words found within these lexicons. Based on the final computed score, each tweet was systematically categorized: instances yielding a total polarity score of zero or higher were labeled as 'Positive', while those with a score strictly below zero were classified as 'Negative'. Furthermore, any duplicate entries were systematically removed during this phase to maintain dataset integrity and uniqueness of the dataset before applying the feature extraction and classification algorithms [18], [19].

D. Data Splitting

Following the data labeling process, the dataset was partitioned into two subsets: a training set and a testing set. This separation was performed prior to feature extraction and data balancing procedures to reduce the risk of data leakage and to support a more reliable evaluation process [20]. The dataset was divided using an 80:20 ratio, a commonly adopted practice in machine learning that provides sufficient data for model training while retaining an adequate portion for validation [21]. Allocating 80% of the data for training and reserving 20% for testing allows the model’s generalization capability to be assessed on previously unseen data, thereby helping to mitigate overfitting [22]. Furthermore, a stratified random sampling technique was applied during this division to maintain the original proportion of positive and negative sentiment classes across both subsets, ensuring consistency with the overall dataset distribution.

E. Feature Extraction

Following the data splitting phase, the textual information underwent a transformation into numerical formats suitable for analysis by the Random Forest algorithm. This transformation was implemented using the Term Frequency–Inverse Document Frequency (TF-IDF) technique. TF-IDF allocates a numerical value to each word, determined by calculating how often the word appears in a single document (TF) and contrasting it with its occurrence rate across all documents (IDF) [23]. To strictly avoid data leakage, the TF-IDF vectorizer was fitted exclusively on the training set to learn the vocabulary parameters. Subsequently, this established vocabulary was used to transform both the training and unseen testing sets into numerical vectors. This approach works well to emphasize significant and distinguishing words that are helpful for determining sentiment.

Specifically, the TF-IDF vectorizer was configured with a minimum document frequency (min_df) of 4 to ignore highly infrequent or misspelled words, and a maximum document frequency (max_df) of 0.8 to filter out overly common terms. Additionally, a standard Indonesian stop words lexicon from the NLTK library was applied to eliminate meaningless conjunctions and prepositions, thereby enhancing the relevance of the extracted features.

F. Random Oversampling

Following the feature extraction stage, an examination of the training data indicated a noticeable imbalance between positive and negative sentiment instances. To mitigate this imbalance and reduce the potential for majority-class bias, the Random Oversampling (ROS) technique was applied exclusively to the training set [24]. ROS is an oversampling method that randomly duplicates existing instances from the minority class until a balanced class distribution is achieved [25]. Because the technique replicates already labeled feature vectors without modifying their textual or semantic content, the duplicated instances retain the same sentiment labels as their original counterparts. Applying ROS only to the training data helps improve the model’s ability to detect minority-class instances while reducing the risk of data leakage during model evaluation [26].

G. Modelling

Classification modeling was executed employing the Random Forest approach, which was specifically parameterized with 100 decision trees (n_estimators = 100) and a fixed random state (random_state = 42) to ensure a robust and reproducible ensemble learning process. This approach is a method of ensemble learning created using multiple decision trees that are produced from data and feature subsets picked at random [27]. The ultimate class forecast is established via a method of voting where the majority determines the result across all the trees made [28]. The selection of n_estimators = 100 was based on established literature indicating that an ensemble size of approximately 100 trees provides a balanced trade-off between computational efficiency and model stability, particularly in text classification tasks with moderate dataset sizes [29], [30]. Increasing the number of trees generally improves model robustness by reducing variance; however, beyond a certain threshold, the performance gain becomes marginal while computational cost increases significantly. Furthermore, preliminary experiments conducted in this study showed that increasing the number of trees beyond this value did not yield significant improvements

in F1-score, while increasing training time. Other hyperparameters were kept at their default configurations, utilizing the Gini impurity criterion to evaluate split quality across the nodes.

H. Evaluation

The concluding phase of this research structure involves assessing the model, with the goal of confirming how trustworthy it is and determining how well the trained Random Forest method can be applied to different data [31]. This assessment was done by analyzing different performance measures that are derived from the Confusion Matrix. Consequently, before moving on to figuring out the values of the evaluation measures, it is vital to first grasp the four key parts that make up the Confusion Matrix [32], which are listed below:

- True Positive (TP)
  Refers to positive sentiment data that are correctly predicted by the model as belonging to the positive class.
- True Negative (TN)
  Refers to negative sentiment data that are correctly predicted by the model as belonging to the negative class.
- False Positive (FP)
  Refers to negative sentiment data that are incorrectly predicted by the model as positive sentiment, commonly referred to as a Type I error.
- False Negative (FN)
  Refers to positive sentiment data that are incorrectly predicted by the model as negative sentiment, commonly referred to as a Type II error.

Based on the previously mentioned elements, the assessment criteria utilized in this research consist of:

- Accuracy
  Accuracy determines the ratio of accurate forecasts (both affirmative and unsupportive) in comparison to the complete set of data entries, presenting a comprehensive overview of how well the model functions (Equation 1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- Precision
  Precision gauges the degree of correctness between the projected outcomes and the real data, revealing the count of instances identified as positive that genuinely correspond to the category of positive sentiment (Equation 2).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

- Recall
  Recall gauges the model's proficiency in recovering pertinent details, focusing on the ratio of genuine positive emotion information accurately pinpointed by the model relative to all positive instances present (Equation 3).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

- F1-Score
  The F1-score represents the harmonic average of precision and recall, delivering a more thorough depiction of how well a model performs, especially when dealing with datasets where classes are unevenly distributed, since it takes into consideration the equilibrium between these two measurements (Equation 4).

$$F1 - Score = 2x\frac{(Precision\ x\ Recall)}{(Precision + Recall)} \quad (4)$$

III. RESULT

This section presents the experimental results of the sentiment classification analysis on public opinions regarding the *Sekolah Rakyat* Program. It discusses the initial data distribution, the impact of class balancing using Random Oversampling (ROS), the qualitative insights obtained from word cloud visualization, and the performance evaluation of the Random Forest model.

A. Data Crawling

The information gathering phase involved employing online data extraction methods, specifically targeting the social media site X and utilizing the phrase "*Sekolah Rakyat*" as the main search term. Information was obtained from the beginning of January 2025 until December of the same year. After finishing this operation, a total of 8,623 initial posts from the social platform were acquired successfully. Subsequently, the unprocessed data that was gathered was saved as a csv file in preparation for continued manipulation and analysis. A demonstration of the preliminary data outcomes can be seen in the illustration labeled as Figure 2.

	conversation_id_str	created_at	favorite_count	full_text
0	1995113205589303705	Sun Nov 30 12:50:30 +0000 2025	0	Sekolah Rakyat hadir untuk seluruh rakyat Indo...
1	1995113115931799932	Sun Nov 30 12:50:08 +0000 2025	0	Sekolah Rakyat membuka akses pendidikan layak ...
2	1994585052274200978	Sun Nov 30 08:20:57 +0000 2025	1	@selputtt Sekolah rakyat adalah investasi bang...
3	1994747534875705427	Sun Nov 30 08:20:48 +0000 2025	1	@pafamore Sekolah rakyat adalah investasi bang...
4	1995045248545034253	Sun Nov 30 08:20:28 +0000 2025	1	Sekolah rakyat adalah investasi bangsa #sekola...

Figure 2. Crawling Data Result

B. Preprocessing

Typically, the unrefined information gathered from social media platforms is disorganized and filled with a lot of unnecessary and unrelated details. Because of this, an extensive data preparation process was performed to refine the

DOI: 10.34148/teknika.v15i1.1446
CC BY SA

data and raise its standard, which made it appropriate for more research, as shown in Table 1.

Table 1. Preprocessing Process

Preprocessing Flow	Before	After
Cleaning	@cheerupawa *Rilll aku habis nonton sidang amdal sekolah rakyat program unggulan nya prabs jadi ini program dikebut bangett. Sekitar 5 ha hutan pinus didaerahku bakal ditebang. Sedihhh baangett semuanya cuman demi ambisi pemerintah*	*Rilll aku habis nonton sidang amdal sekolah rakyat program unggulan nya prabs jadi ini program dikebut bangett Sekitar ha hutan pinus didaerahku bakal ditebang Sedihhh baangett semuanya cuman demi ambisi pemerintah*
Case Folding	*Rilll aku habis nonton sidang amdal sekolah rakyat program unggulan nya prabs jadi ini program dikebut bangett Sekitar ha hutan pinus didaerahku bakal ditebang Sedihhh baangett semuanya cuman demi ambisi pemerintah*	*rilll aku habis nonton sidang amdal sekolah rakyat program unggulan nya prabs jadi ini program dikebut bangett sekitar ha hutan pinus didaerahku bakal ditebang sedihhh baangett semuanya cuman demi ambisi pemerintah*
Normalization	*rilll aku habis nonton sidang amdal sekolah rakyat program unggulan nya prabs jadi ini program dikebut bangett sekitar ha hutan pinus didaerahku bakal ditebang sedihhh baangett semuanya cuman demi ambisi pemerintah*	*riil aku telah menonton sidang amdal sekolah rakyat program unggulannya Prabowo jadi ini program dipercepat sangat sekitar hektare hutan pinus di daerahku akan ditebang sedih sangat semuanya hanya demi ambisi pemerintah*
Tokenization	*riil aku telah menonton sidang amdal sekolah rakyat program unggulannya Prabowo jadi ini program dipercepat sangat sekitar hektare hutan pinus di daerahku akan ditebang sedih sangat semuanya hanya demi ambisi pemerintah*	*['riil', 'aku', 'telah', 'menonton', 'sidang', 'amdal', 'sekolah', 'rakyat', 'program', 'unggulannya', 'Prabowo', 'jadi', 'ini', 'program', 'dipercepat', 'sangat', 'sekitar', 'hektare', 'hutan', 'pinus', 'di', 'daerahku', 'akan', 'ditebang', 'sedih', 'sangat', 'semuanya', 'hanya', 'demi', 'ambisi', 'pemerintah']*
Stopword Removal	*['riil', 'aku', 'telah', 'menonton', 'sidang', 'amdal', 'sekolah', 'rakyat', 'program', 'unggulannya', 'Prabowo', 'jadi', 'ini', 'program', 'dipercepat', 'sangat', 'sekitar', 'hektare', 'hutan', 'pinus', 'di', 'daerahku', 'akan', 'ditebang', 'sedih', 'sangat', 'semuanya', 'hanya', 'demi', 'ambisi', 'pemerintah']*	*['riil', 'menonton', 'sidang', 'amdal', 'sekolah', 'rakyat', 'program', 'unggulannya', 'Prabowo', 'program', 'dipercepat', 'hektare', 'hutan', 'pinus', 'daerahku', 'ditebang', 'sedih', 'ambisi', 'pemerintah']*
Stemming	*['riil', 'menonton', 'sidang', 'amdal', 'sekolah', 'rakyat', 'program', 'unggulannya', 'Prabowo', 'program', 'dipercepat', 'hektare', 'hutan', 'pinus', 'daerahku', 'ditebang', 'sedih', 'ambisi', 'pemerintah']*	*['riil', 'tonton', 'sidang', 'amdal', 'sekolah', 'rakyat', 'program', 'unggul', 'prabowo', 'program', 'cepat', 'hektare', 'hutan', 'pinus', 'daerah', 'tebang', 'sedih', 'ambisi', 'perintah']*

C. Data Labelling

Following the data cleanup, the lexicon-based labeling process was executed to determine the ground-truth sentiment of each tweet. Using the Indonesian sentiment dictionaries, polarity scores were calculated for every preprocessed text. Tweets yielding a total polarity score of zero or higher were categorized into the positive sentiment group, whereas those

with a score below zero were assigned to the negative sentiment group.

The preliminary distribution of the generated sentiment labels revealed a clear class imbalance as shown in Figure 3. Out of a total of 8,623 data instances, positive sentiment constituted 5,895 tweets (69.4%), while negative sentiment accounted for 2,638 tweets (30.6%). This unequal distribution poses a technical challenge, as classification models trained on imbalanced data tend to favor the majority class, potentially reducing the model's sensitivity toward minority sentiment detection.

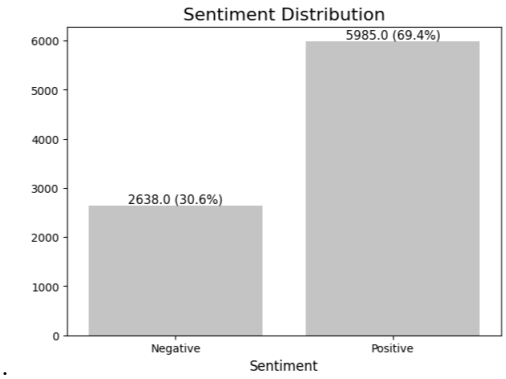


Figure 3. Data Labelling Result

D. Data Splitting

Following the labeling process, the dataset was partitioned into training and testing subsets using a stratified 80:20 ratio to preserve the original sentiment distribution across both subsets. From the total of 8,623 tweets, 6,898 instances were allocated to the training set, while 1,725 instances were reserved for the testing set (Figure 4). The stratified sampling approach ensured that the proportion of positive and negative sentiment classes remained consistent with the overall dataset distribution.

This separation was performed prior to feature extraction and data balancing procedures to prevent data leakage and ensure an unbiased evaluation of the classification model. By isolating the testing set at an early stage, the model's performance was assessed on genuinely unseen data, thereby providing a reliable measure of its generalization capability.

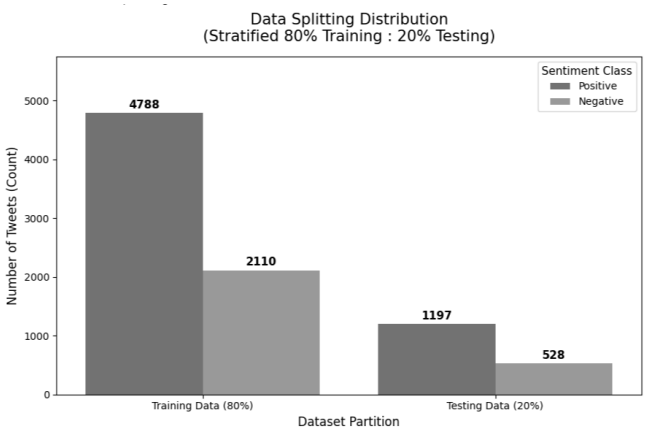


Figure 4. Data Splitting Distribution

E. Feature Extraction

Following the dataset preparation, the textual data was transformed into numerical vectors employing the Term Frequency–Inverse Document Frequency (TF-IDF) technique. By assigning higher weights to words that are highly significant to specific sentiments while penalizing overly common terms, this feature extraction step effectively surfaced the most distinguishing features within the dataset.

According to the feature extraction outcomes, the most prominent words associated with the positive sentiment group are "*didik*," "*program*," "*rakyat*," "*sekolah*," "*gratis*," "*bangun*," and "*presiden*" (Figure 5). The frequent appearance of the words "gratis" and "sekolah" indicates the public's positive perception of the initiative's effectiveness in delivering reasonably priced educational opportunities. Furthermore, the words "*bangun*" and "*didik*" demonstrate popular backing for governmental projects aimed at growing infrastructure and boosting educational standards for the populace.



Figure 5. Positive Sentiment Wordcloud

In the negative sentiment class, the most prominent word features include "*sekolah*", "*rakyat*", "*miskin*", "*anak*", "*bayar*", and "*butuh*" (Figure 6). The dominance of the terms "miskin" and "anak" indicates public concerns related to social inequality that affects access to education for underprivileged communities. In addition, the presence of the terms "*bayar*" and "*butuh*" highlights issues related to unexpected costs or facilities that are still perceived as inadequate in practice.



Figure 6. Negative Sentiment Wordcloud

F. Random Oversampling

To resolve the significant class imbalance observed during the data labeling phase, the Random Oversampling (ROS) method was implemented. The outcome of this balancing procedure is presented in Table 2, illustrating the equalization of the minority negative sentiment class to match the majority positive sentiment class.

DOI: 10.34148/teknika.v15i1.1446
TEKNIKA, Volume 15(1), March 2026, pp. 168-177
ISSN 2549-8037, EISSN 2549-8045

To address methodological concerns regarding the integrity of the text and its corresponding sentiment after sampling, it is important to clarify the operational mechanism of Random Oversampling (ROS). This technique works by randomly duplicating existing instances from the minority class that have already been labeled during the annotation process. It does not generate synthetic text, modify semantic structures, or alter the predefined lexicon-based polarity scores. As a result, the duplicated tweets retain the same sentiment labels as their original instances. This approach helps preserve label consistency and reduces the risk of unintended misclassification during the balancing process.

Table 2. Balancing Result

Sentiment	Before Balancing	After Balancing
Negative	2,110	4,788
Positive	4,788	4,788

G. Evaluation

The Random Forest model's effectiveness was assessed utilizing a test dataset that represented 20% of the complete labeled dataset, which equated to 1,725 samples. The evaluation of performance relied on the Confusion Matrix, a tool that juxtaposes the forecasts made by the model against the true, known labels (Figure 7).

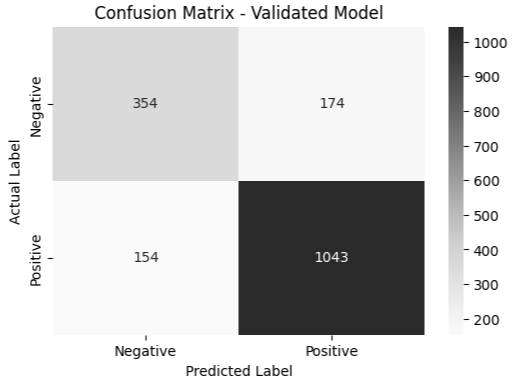


Figure 7. Confusion Matrix Result

Table 3. Model Performance

	Precision	Recall	F1-Score	Support
negative	0.70	0.67	0.68	528
positive	0.86	0.87	0.86	1,197
accuracy			0.81	1,725
macro avg	0.78	0.77	0.77	1,725
weighted avg	0.81	0.81	0.81	1,725

The model's performance assessment (Table 3) revealed that it accurately categorized 1,397 cases (both correctly identified positives and negatives) out of a total of 1,725 testing instances, alongside a limited set of incorrect categorizations. A summarized breakdown of the model's performance metrics is shown below:

- Accuracy, the model demonstrated an 80.99% accuracy rate. This indicates that the model demonstrates stable classification performance on unseen Indonesian social media data.
- Precision, the weighted average precision score was 0.81, which means that the model is suitably accurate when forecasting sentiment categories without overwhelmingly biasing the predictions toward the majority class.
- Recall, the weighted average recall score, also at 0.81, points to the model's proficiency in correctly identifying the majority of pertinent instances from the actual data points, maintaining a fair detection rate even for the minority negative class.
- F1-Score, the weighted average F1-score, recorded at 0.81, shows a solid equilibrium between precision and recall, thus verifying how well the combination of the Random Forest algorithm and Random Oversampling works in processing imbalanced, sentiment-based data.

In general, these outcomes make it clear that the suggested method works well to accurately chart public sentiment, where positive sentiment (accounting for approximately 69.4% of the testing data) is significantly more widespread than negative sentiment. This observation implies that the public generally has a good impression of the *Sekolah Rakyat* Program.

IV. DISCUSSION

To establish a clear empirical foundation for evaluating the proposed approach, a comparative analysis between the Multinomial Naïve Bayes (NB) baseline and the proposed Random Forest (RF) model was conducted. Both models were implemented using the same preprocessing pipeline, TF-IDF configuration, stratified 80:20 data split, and evaluation metrics to ensure experimental fairness.

Table 4. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Multinomial Naïve Bayes	0.7629	0.79	0.76	0.77
Random Forest	0.8099	0.81	0.81	0.81

As presented in Table 4, the proposed Random Forest model consistently outperforms the Naïve Bayes baseline across all evaluation metrics. The overall accuracy improved by approximately 4.7 percentage points, while the weighted F1-score increased from 0.77 to 0.81. This improvement indicates that the ensemble-based Random Forest approach provides more balanced predictive performance, particularly in handling minority class instances.

The superior performance of Random Forest can be explained by its ability to model complex feature interactions within high-dimensional TF-IDF representations. Unlike Naïve Bayes, which assumes conditional independence among features, Random Forest constructs multiple decision trees and

aggregates their outputs through majority voting. This mechanism enables the model to better capture non-linear relationships and contextual dependencies within Indonesian social media text, which often contains informal expressions, abbreviations, and lexical variations.

Another critical factor influencing the results is the handling of class imbalance. The original dataset was dominated by positive sentiment (approximately 69.4%), creating the potential for majority-class bias. To address this issue, Random Oversampling (ROS) was applied exclusively to the training dataset, thereby increasing the representation of minority class instances without modifying their original semantic content. The resulting performance, particularly the negative class recall of 0.67, suggests that the model achieved improved sensitivity toward minority sentiment without significantly degrading overall classification stability. This finding aligns with previous studies indicating that oversampling strategies can enhance minority class detection while maintaining acceptable generalization performance [11], [12].

The effectiveness of the proposed approach is also supported by the structured preprocessing pipeline. Indonesian social media text is characterized by slang, informal spelling, and morphological variation. Through normalization, stemming, and stopword removal, the textual data were standardized prior to feature extraction. The application of TF-IDF weighting further enhanced discriminative capability by emphasizing sentiment-relevant terms while reducing the influence of overly common words. The combination of preprocessing and ensemble learning contributes to the robustness of the classification model in handling sparse and high-dimensional feature spaces.

When compared with previous research on public policy sentiment analysis [1], [2] the findings of this study demonstrate consistent sentiment distribution patterns, where one dominant sentiment class tends to prevail in online discourse. However, this study extends prior work by explicitly incorporating imbalance handling and baseline comparison within the experimental framework. This methodological enhancement strengthens the validity of performance claims and provides a more rigorous evaluation of classification effectiveness.

Beyond quantitative evaluation, qualitative insights derived from word cloud visualization further contextualize the results. Frequently occurring terms in the positive sentiment class, such as “gratis,” “sekolah,” and “didik,” reflect public appreciation for increased educational access and infrastructure development. Conversely, terms such as “miskin,” “bayar,” and “anak” in the negative sentiment class indicate concerns related to socioeconomic implications and policy implementation challenges. These patterns suggest that public discourse surrounding the *Sekolah Rakyat* Program contains both supportive and critical perspectives, highlighting the multidimensional nature of public opinion [3], [6].

It is important to note that this study adopts a binary sentiment classification framework consisting only of positive and negative classes, without incorporating a neutral category. While this approach simplifies the modeling process and may improve classification performance, it may also limit the representational richness of public opinion. In real-world public policy discourse, neutral sentiments often reflect informational, ambiguous, or undecided opinions that do not explicitly convey polarity. The exclusion of this category may lead to forced classification, where inherently neutral expressions are assigned to either positive or negative classes, potentially affecting overall model accuracy and interpretability. Therefore, future research should consider multi-class sentiment classification to better capture the complexity of public responses and provide a more nuanced evaluation of policy perception.

Overall, the integration of Random Forest, TF-IDF feature extraction, and Random Oversampling demonstrates an empirically supported framework for sentiment analysis in imbalanced Indonesian social media datasets. By incorporating baseline comparison and imbalance mitigation strategies, this study provides a more comprehensive and methodologically sound approach to evaluating public policy sentiment through machine learning techniques.

V. CONCLUSION

This study investigated public sentiment toward the *Sekolah Rakyat* Program on social media platform X using a machine learning framework integrating TF-IDF feature extraction, Random Oversampling (ROS), and the Random Forest classifier. The results demonstrate that handling class imbalance and employing an ensemble-based learning approach contribute to improved sentiment classification performance on Indonesian social media data.

The proposed Random Forest model achieved an accuracy of 80.99% with weighted precision, recall, and F1-score values of 0.81 on the unseen testing dataset. When compared to the Multinomial Naïve Bayes baseline model, which obtained an accuracy of 76.29%, the proposed approach showed measurable performance improvement. This comparison confirms that the ensemble-based Random Forest method is more effective in capturing complex lexical patterns within high-dimensional TF-IDF representations than traditional probabilistic classifiers.

The application of Random Oversampling exclusively to the training data helped mitigate majority-class bias without introducing data leakage, thereby improving minority class detection while maintaining overall model stability. These findings highlight the importance of imbalance handling strategies in sentiment analysis of real-world social media datasets.

From a substantive perspective, the sentiment distribution indicates that public opinion toward the *Sekolah Rakyat* Program is predominantly positive, accounting for approximately 69–70% of the dataset. However, the presence of critical sentiments related to socioeconomic concerns suggests that public discourse remains multidimensional. Therefore, sentiment analysis can serve as a complementary evaluative tool for policymakers in assessing public reception and identifying areas requiring further attention.

DOI: 10.34148/teknika.v15i1.1446

Future research may expand this study by incorporating more advanced deep learning architectures, such as Long Short-Term Memory (LSTM) networks or transformer-based language models specifically trained on Indonesian corpora, to further enhance contextual understanding of sentiment expressions. Additionally, future work could explore multi-class sentiment categorization (e.g., positive, negative, and neutral) and conduct cross-validation or longitudinal analysis to examine temporal sentiment dynamics. Expanding the dataset to include data from multiple social media platforms may also improve generalizability and provide a broader representation of public opinion. Furthermore, investigating alternative imbalance handling techniques, such as SMOTE or cost-sensitive learning, may offer additional insights into optimizing classification performance in imbalanced policy-related datasets.

REFERENCES

[1] N. Hadi and D. Sugiarto, "Analisis Sentimen Pembangunan IKN pada Media Sosial X Menggunakan Algoritma SVM, Logistic Regression dan Naïve Bayes," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 10, no. 1, pp. 37–49, Jan. 2025, doi: 10.30591/jpit.v10i1.7106.

[2] W. B. Zulfikar, A. R. Atmadja, and S. F. Pratama, "Sentiment Analysis on Social Media Against Public Policy Using Multinomial Naive Bayes," *Scientific Journal of Informatics*, vol. 10, no. 1, pp. 25–34, Jan. 2023, doi: 10.15294/sji.v10i1.39952.

[3] S. N. Russell, L. Rao-Graham, and M. McNaughton, "Mining social media data to inform public health policies: a sentiment analysis case study," *Revista Panamericana de Salud Publica/Pan American Journal of Public Health*, vol. 48, 2024, doi: 10.26633/RPSP.2024.79.

[4] "Mensos Lantik 860 Guru *Sekolah Rakyat* - Tribatanews Polri.". Dec. 14, 2025. [Online]. Available: https://tribratanews.polri.go.id/blog/nasional-3/mensos-lantik-860-guru-sekolah-rakyat-95697

[5] "*Sekolah Rakyat*: Gebrakan Satu Tahun Pemerintahan Presiden Prabowo Subianto untuk Memutus Rantai Kemiskinan - Media Keuangan.". Dec. 14, 2025. [Online]. Available: https://mediakeuangan.kemenkeu.go.id/article/show/sekolah-rakyat-gebrakan-satu-tahun-pemerintahan-presiden-prabowo-subianto-untuk-memutus-rantai-kemiskinan

[6] R. R. Fitri, A. -, and W. Apriandari, "Penggunaan Random Forest Dalam Sistem Klasifikasi Kecemasan Pada Generasi Z," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 3, Jul. 2025, doi: 10.23960/jitet.v13i3.6905.

[7] M. S. Kraiem, F. Sánchez-Hernández, and M. N. Moreno-García, "Selecting the suitable resampling strategy for imbalanced data classification regarding dataset properties. An approach based on association models," *Applied Sciences (Switzerland)*, vol. 11, no. 18, Sep. 2021, doi: 10.3390/app11188546.

[8] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Information (Switzerland)*, vol. 14, no. 1, Jan. 2023, doi: 10.3390/info14010054.

[9] D. Triyana, M. Muharrom, A. Haromainy, and H. Maulana, "Implementasi Metode Ensemble Majority Vote Pada Algoritma Naive Bayes Dan Random Forest Untuk Analisis Sentimen Twitter Harga Tiket Pesawat Domestik," *Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 4, 2024, doi: 10.36040/jati.v8i4.10475.

[10] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, Dec. 2024, doi: 10.58496/BJML/2024/007.

[11] A. H. Putra and A. Salam, "A Comparative Performance of SMOTE, ADASYN and Random Oversampling in Machine Learning Models on Prostate Cancer Dataset," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 3, p. 603, 2025, doi: 10.30871/jaic.v9i3.9308.

[12] S. Diantika, H. Nalatissifa, N. Maulidah, R. Supriyadi, and A. Fauzi, "Penerapan Teknik Random Oversampling Untuk Memprediksi Ketepatan Waktu Lulus Menggunakan Algoritma Random Forest," *Computer Science (CO-SCIENCE)*, vol. 4, no. 1, pp. 11–18, Jan. 2024, doi: 10.31294/coscience.v4i1.1996.

[13] N. A. Semary, W. Ahmed, K. Amin, P. Pławiak, and M. Hammad, "Enhancing machine learning-based sentiment analysis through feature extraction techniques," *PLoS One*, vol. 19, no. 2 February, Feb. 2024, doi: 10.1371/journal.pone.0294968.

[14] A. Anies and M. Ikhsan, "Sentiment Analysis of 'Free Lunch for Children' Policy on Social Media X Using Random Forest Algorithm," *Journal of Information Systems and Informatics*, vol. 7, no. 1, pp. 649–662, Mar. 2025, doi: 10.51519/journalisi.v7i1.1039.

[15] M. Alif, M. Alamsyah, and M. F. Arif, "Analisis Sentimen Twitter Tentang Pinjaman Online di Indonesia Menggunakan Metode Random Forest," *Jurnal Ilmiah Teknik Informatika Dan Sistem Informasi*, vol. 13, no. 2, p. 1410, Aug. 2024, doi: 10.35889/jutisi.v13i2.2215.

[16] S. J. Pipin and H. Kurniawan, "Analisis Sentimen Kebijakan MBKM Berdasarkan Opini Masyarakat di Twitter Menggunakan LSTM," *Jurnal Sifo Mikroskil*, vol. 23, pp. 1–5, doi: 10.55601/jsm.v23i2.900.

[17] A. Syah, F. Nurdiyansyah, and A. Y. Rahman, "Analisis Sentimen Aplikasi Shopee, Tokopedia, Lazada dan Blibli Menggunakan Leksikon Dan Random Forest," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3S1, Oct. 2024, doi: 10.23960/jitet.v12i3s1.5155.

[18] X. Zhang, "Text classification algorithms exploration on sentiment analysis," *Applied and Computational Engineering*, vol. 5, no. 1, pp. 99–103, May 2023, doi: 10.54254/2755-2721/5/20230541.

[19] W. Xu, J. Chen, Z. Ding, and J. Wang, "Text Sentiment Analysis and Classification Based on Bidirectional

Gated Recurrent Units (GRUs) Model." doi: 10.54254/2755-2721/77/20240670.

[20] A. Apicella, F. Isgrò, and R. Prevete, "Don't push the button! Exploring data leakage risks in machine learning and transfer learning," *Artif. Intell. Rev.*, vol. 58, no. 11, Nov. 2025, doi: 10.1007/s10462-025-11326-3.

[21] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, "Trade-off between training and testing ratio in machine learning for medical image processing," *PeerJ Comput. Sci.*, vol. 10, 2024, doi: 10.7717/PEERJ-CS.2245.

[22] S. Kapoor and A. Narayanan, "Leakage and the Reproducibility Crisis in ML-based Science," Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.07048

[23] M. D. Rizkiyanto, M. D. Purbolaksono, and W. Astuti, "Sentiment Analysis Classification on PLN Mobile Application Reviews using Random Forest Method and TF-IDF Feature Extraction," *INTEK: Jurnal Penelitian*, vol. 11, no. 1, pp. 37–43, Apr. 2024, doi: 10.31963/intek.v11i1.4774.

[24] G. R. Ashisha, X. A. Mary, E. G. M. Kanaga, J. Andrew, and R. J. Eunice, "Random Oversampling-Based Diabetes Classification via Machine Learning Algorithms," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, Dec. 2024, doi: 10.1007/s44196-024-00678-3.

[25] C. Yang, E. A. Fridgeirsson, J. A. Kors, J. M. Reps, and P. R. Rijnbeek, "Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data," *J. Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-023-00857-7.

[26] A. Moreo, A. Esuli, and F. Sebastiani, "Distributional random oversampling for imbalanced text classification," in *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, Inc, Jul. 2016, pp. 805–808. doi: 10.1145/2911451.2914722.

[27] M. A. Ganaie, M. Tanveer, P. N. Suganthan, and V. Snasel, "Oblique and rotation double random forest," Aug. 2022, doi: 10.1016/j.neunet.2022.06.012.

[28] P. Antonio, E. Lim, and C. H. Park, "A Collaborative Ensemble Construction Method for Federated Random Forest," 2024. doi: 10.1016/j.eswa.2024.124742.

[29] Y. A. Seraphina and P. H. Gunawan, "Sentiment Analysis of Public Responses Regarding The Use of Electric Cars in Indonesia with Support Vector Machine and Random Forest Methods," *The Indonesian Journal of Computer Science*, vol. 14, no. 1, Feb. 2025, doi: 10.33022/ijcs.v14i1.4649.

[30] I. A. Nur Sabrina, T. Triastuti Wuryandari, and R. Dapa, "Implementation of Random Forest Algorithm in Classifying Public Sentiment Towards Free Nutritious Meal Program," *Current Science Research Bulletin*, vol. 02, no. 05, May 2025, doi: 10.55677/csrb/02-V02I05Y2025.

[31] W. Zhuo and A. Ahmad, "HCRF: an improved random forest algorithm based on hierarchical clustering," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 38, no. 1, p. 578, Apr. 2025, doi: 10.11591/ijeecs.v38.i1.pp578-586.

[32] J. M. Baladjay, N. Riva, L. A. Santos, D. M. Cortez, C. Centeno, and A. A. R. Sison, "Performance evaluation of random forest algorithm for automating classification of mathematics question items," *World Journal of Advanced Research and Reviews*, vol. 18, no. 2, pp. 034–043, May 2023, doi: 10.30574/wjarr.2023.18.2.0762.