Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3087~3096
e-ISSN: 2548-6861

Sentiment Analysis of the TPKS Law on Twitter: A Comparative Study of Classification Algorithm Performance

Heni Sapta Mawar [1]*, Majid Rahardi [2]*
*Informatics, Universitas Amikom Yogyakarta
henisaptamawar@students.amikom.ac.id [1], majid@amikom.ac.id [2]

Article Info

Article history:

Received 2025-10-15
Revised 2025-10-31
Accepted 2025-11-05

Keyword:
Machine Learning,
Random Forest,
Sentiment Analysis,
Support Vector Machine,
TF-IDF.

ABSTRACT

The enactment of Law Number 12 of 2022 concerning the Crime of Sexual Violence (UU TPKS) has generated substantial public discussion on social media, particularly on Twitter. This study compares the performance of six machine learning algorithms—Naive Bayes, K-Nearest Neighbors, Logistic Regression, Support Vector Machine, Random Forest, and XGBoost—for sentiment analysis of Indonesian-language tweets related to the UU TPKS. A total of 2,351 tweets collected between 2022 and 2025 were preprocessed through case folding, normalization, stopword removal, and stemming using the Sastrawi library, then transformed into numerical features using TF-IDF. Exploratory data analysis revealed 18,230 features, indicating a high-dimensional and sparse dataset. Model evaluation employed 10-Fold Cross-Validation with class weighting to address class imbalance (80.22% negative vs. 19.78% positive). Results show that Support Vector Machine and Random Forest achieved the best performance, with accuracy of 85.35% and F1-score of 0.83. ROC-AUC values of 0.92 and 0.91 confirmed their discriminative strength, and a paired t-test indicated their superiority was statistically significant ($p < 0.05$). Both models also demonstrated robust computational efficiency and generalization. This study highlights the effectiveness of combining TF-IDF with margin-based and ensemble methods for Indonesian-language sentiment analysis and suggests future enhancements using transformer-based models such as IndoBERT or IndoBERTweet for deeper semantic understanding.

I. INTRODUCTION

Sexual violence remains a serious, unresolved problem in Indonesia. The high number of cases reported each year indicates a deficit in social and legal protection for victims [36]–[38]. In an effort to strengthen this protection, the Indonesian government passed Law Number 12 of 2022 concerning Criminal Acts of Sexual Violence (UU TPKS). This regulation marks a significant milestone in victim-centered legal reform, encompassing prevention, treatment, and recovery for victims of sexual violence [17].

However, the enactment of the TPKS Law sparked widespread and polarized public discourse, particularly on social media platforms such as Twitter [36]. This platform serves as a primary platform for people to express their opinions, support, and criticism of the policy. The large volume of text data generated reflects the complexity of public perception, which is not only emotionally diverse but also deeply embedded in social, cultural, and political contexts. To systematically understand the dynamics of public opinion, a machine learning-based analytical approach through sentiment analysis is needed, which has proven effective in various public policy contexts [7], [22], [23].

Sentiment analysis using machine learning algorithms allows researchers to automatically classify public opinion texts into positive, negative, or neutral categories [8], [25]. This approach has been applied in various studies in Indonesia, such as analyses of the Omnibus Law and PPKM policies during the COVID-19 pandemic [7], [22]. However, previous studies have shown varying results in determining the most optimal classification algorithm. Some studies report that Support Vector Machines (SVMs) produce the highest accuracy on text classification tasks [9], [24], [35], while others find that ensemble-based models such as Random

Forest (RF) or XGBoost provide more stable results and are resistant to overfitting [17], [19], [27], [31].

Recent studies further emphasize this diversity of findings. Al Faridzi et al. (2023) [9] demonstrated the robustness of SVM in handling imbalanced sentiment data, while Agustine & Jayadi (2025) [12] found that Random Forest outperformed other algorithms in multi-domain sentiment analysis. Similarly, Nurdiyanti & Utami (2024) [17] showed that model performance may vary significantly depending on the thematic domain—such as sensitive social and legal topics—highlighting the importance of domain-specific evaluation in sentiment analysis.

This variation in results indicates that the performance of classification algorithms is highly dependent on the characteristics of the dataset and the complexity of the topic being analyzed. This phenomenon aligns with the No Free Lunch Theorem, which states that no single algorithm is consistently superior for all types of problems [16]. In the context of the TPKS Law, public discourse on social media is highly complex—a mixture of formal and informal language, with emotional expressions and specific legal terms. This makes research findings in other domains not necessarily directly applicable [11], [21].

Based on this research gap, this study aims to conduct a comparative analysis of the performance of classification algorithms in sentiment analysis of the TPKS Law on Twitter, focusing on six machine learning models: Naive Bayes (NB), K-Nearest Neighbors (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and XGBoost. This study uses the Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction method, which has been widely used in Indonesian text analysis [7], [17], [34], [42]. Through this approach, the research is expected to identify the most accurate and efficient algorithm for measuring public perception of government policies, while also providing an empirical contribution to the use of machine learning in policy informatics and computational social science in Indonesia.

II. METHODS

The research methodology is illustrated in Figure 1 and consists of five main stages: (1) data collection, (2) text pre-processing, (3) feature extraction using TF-IDF, (4) modeling, and (5) performance evaluation. This study conducts a comparative analysis of six machine learning classification algorithms: Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), XGBoost, K-Nearest Neighbors (KNN), and Logistic Regression (LR).

After the data preprocessing stage, the text data was transformed into numeric vector representations using the Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction method. For model evaluation, the dataset was partitioned into 80% training data and 20% test data. Model validation on the training data was performed using the K-Fold Cross-Validation technique to ensure stability and reliability.

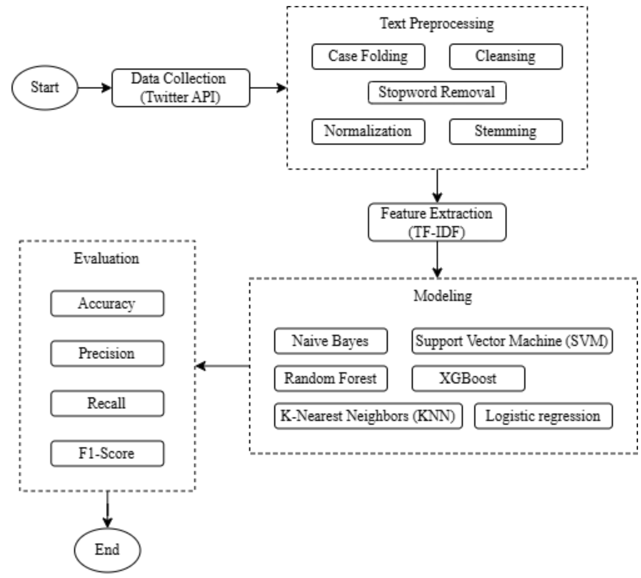

Figure 1. Research Flowchart

The modeling stage involved a comparative analysis of six classification algorithms: Naive Bayes, Support Vector Machine (SVM), Random Forest, XGBoost, K-Nearest Neighbors (KNN), and Logistic Regression as the baseline model. Hyperparameter optimization for each model was performed using a grid search methodology within a cross-validation framework, with F1-Score as the primary metric to address potential class imbalance. Finally, the final performance of each optimized model was measured on the test data, where key metrics such as Accuracy, Precision, Recall, and F1-Score were extracted from the Confusion Matrix.

A. Data Collection

The data used in this study consists of textual public opinion collected from the social media platform Twitter. The collection process was carried out using a crawling technique through Twitter's official Application Programming Interface (API). Keywords such as "UU TPKS", "RUU PKS", and "sexual violence" were used to capture relevant data.

The data collection was conducted during the 2022–2025 period, covering the timeline from the enactment of the Law on the Crime of Sexual Violence (UU TPKS) to the ongoing public debates regarding its implementation. This time frame was chosen to capture both immediate and evolving public sentiment surrounding the policy. The crawling process was executed using manual scripts based on Twitter's official API authentication, ensuring compliance with the platform's data usage policies. Several keywords—including *"UU TPKS"*, *"RUU PKS"*, and *"sexual violence"*—were used to identify and extract relevant posts.

This process initially generated 2,389 raw data points. After a cleaning process to remove empty and duplicate data, the final dataset consisted of 2,351 tweets. Each collected tweet then underwent a manual labeling process by several annotators to assign a sentiment label (positive or negative). This process aims to produce high-quality ground truth that serves as a basis for training and testing the model.

This study strictly adheres to Twitter's data usage policies and ethical standards for public data research. All data collected were publicly available tweets and did not include any private or personally identifiable information such as usernames, profile photos, or user IDs. Before analysis, all tweet identifiers were anonymized to ensure user privacy. The data collection process was conducted using an officially authenticated API token in compliance with Twitter's Terms of Service. Furthermore, the research follows ethical principles of digital and social media research, ensuring that the data are used solely for academic purposes without any attempt to trace or identify individual users.

B. *Text Preprocessing*

The labeled raw data then undergoes several preprocessing stages to clean and standardize the text. This process is essential to address the unique challenges of Indonesian-language social media text, which often contains informal expressions, slang, abbreviations, misspellings, and code-mixing between Indonesian and English. Without thorough normalization and cleaning, the resulting feature space would become large and sparse, potentially degrading the performance of classification algorithms.

The preprocessing stages carried out in this study include the following steps:

- Case Folding: Converting all text in the corpus to lowercase to ensure consistency during processing.
- Cleansing: Removing irrelevant elements such as URLs, mentions (@username), hashtags (#), numbers, and other non-alphanumeric characters that may interfere with analysis and contribute noise to the dataset.
- Normalization: Converting non-standard or slang words into standard Indonesian using a predefined normalization dictionary, which helps standardize lexical variations commonly found in online discourse.
- Stopword Removal: Eliminating common words that do not carry significant sentiment weight such as "yang," "di," "dan," and "adalah" using an Indonesian stopword list to improve the representativeness of meaningful terms.
- Stemming: Reducing affixed or derived words to their root form (e.g., "menghasil" becomes "hasil") to minimize word variation and reduce feature dimensionality. This process was performed using the Sastrawi stemming library, which is specifically optimized for Bahasa Indonesia.

In addition to the normalization and stemming stages, this study also encountered several linguistic challenges specific to the Indonesian language as used on social media. Many tweets contain non-standard words, mixed-language expressions, as well as sarcasm or irony, which are difficult to interpret literally using statistical models such as TF-IDF. Although the normalization process helps reduce lexical variation, sarcastic or ambiguous expressions remain a major challenge for sentiment classification. This complexity highlights that linguistic aspects significantly affect model accuracy, particularly within Indonesia's unique socio-cultural context.

Therefore, the findings of this study should be interpreted with these limitations in mind. Future research may build upon this work by applying contextual embedding approaches, such as FastText or IndoBERT, to capture deeper semantic meaning and emotional nuance from social media text.

As a continuation of these preprocessing stages, this study then performed feature extraction using the Term Frequency–Inverse Document Frequency (TF-IDF) method to transform text into numerical representations for the modeling process.

C. *Feature Extraction*

After the text is cleaned, the next step is to convert it into a numerical vector representation using the Term Frequency–Inverse Document Frequency (TF-IDF) weighting method. This technique evaluates how important a word is within a document compared to its occurrence across the entire dataset. The Term Frequency (TF) represents how often a term appears in a document, while the Inverse Document Frequency (IDF) reflects the uniqueness or rarity of that term across all documents. By combining these two measures, TF-IDF emphasizes words that are frequent in a specific document but less common in others, allowing the model to focus on terms that carry the most meaningful information for classification.

D. *Classification Algorithm*

This study systematically compares six classification algorithms to address the diverse and sometimes conflicting findings in the literature regarding predictive models for text data. The evaluated algorithms include Naive Bayes (NB), a probabilistic classifier [35]; K-Nearest Neighbors (KNN), an instance-based non-parametric algorithm [5], [18]; Logistic Regression (LR), a robust linear model [15], [24]; Support Vector Machine (SVM), which is effective in high-dimensional spaces [2], [9], [24]; Random Forest (RF), an ensemble method that is robust against overfitting [19], [31]; and XGBoost, an efficient implementation of gradient boosting [27], [8]. To ensure valid evaluation and prevent overfitting, the dataset was partitioned into 80% training data and 20% testing data. Furthermore, a 10-Fold Cross-Validation technique was applied to the training data to ensure model stability and independence from specific data partitions [20], [34].

The selection of these six algorithms is theoretically grounded on their differing learning paradigms, which allow for a comprehensive comparative analysis across probabilistic, margin-based, and ensemble approaches. Naive Bayes (NB) represents the probabilistic family, relying on Bayes' theorem with conditional independence assumptions. K-Nearest Neighbors (KNN) is an instance-based algorithm that classifies based on distance metrics, reflecting local data structures. Logistic Regression (LR) is a linear discriminative model that estimates class probabilities through a sigmoid function. Support Vector Machine (SVM) belongs to the margin-based category, which seeks an optimal separating hyperplane in high-dimensional space, making it robust to sparse TF-IDF features. Random Forest (RF) and XGBoost, on the other hand, represent ensemble-based learning approaches that combine multiple decision trees to improve generalization and reduce overfitting.

By integrating algorithms with distinct theoretical foundations, this study aims to ensure a fair and diverse evaluation of classification performance in sentiment analysis of Indonesian-language text, consistent with the *No Free Lunch Theorem*, which posits that no single algorithm performs best across all data domains.

To ensure a fair comparison, Logistic Regression was designated as the baseline model due to its computational efficiency and clear coefficient interpretation [15], [24]. This model serves as a reference point for evaluating the performance gains of more complex classifiers such as SVM, Random Forest, and XGBoost.

Comparative studies in Indonesian-language sentiment analysis, including Sukma et al. (2020) on the Omnibus Law, Agustiningsih et al. (2022) on COVID-19 vaccine sentiment, and Ansori & Holle (2022) on Twitter-based public opinion, provide relevant empirical benchmarks. These studies collectively highlight that while linear models such as Logistic Regression establish strong baselines, ensemble and margin-based methods often yield superior generalization on high-dimensional text data.

In this study, hyperparameter optimization for each algorithm was conducted using a grid search approach combined with cross-validation to obtain the best configuration while preventing overfitting. The optimal settings obtained were as follows: K = 5 with Euclidean distance for K-Nearest Neighbors (KNN); linear kernel and C = 1 for Support Vector Machine (SVM); 100 trees and max_depth = 10 for Random Forest (RF); learning rate = 0.1 and max_depth = 6 for XGBoost; C = 1 and solver = 'liblinear' for Logistic Regression (LR); and alpha = 1.0 for Naive Bayes (NB). These configurations were selected based on their balance between model complexity, generalization ability, and computational efficiency.

1) *Logistic regression*

Chosen as the baseline model due to its computational efficiency and clear interpretation of coefficients [15], this model estimates the probability of the target class using a logistic function, assuming a linear relationship between features and log-odds [24].

2) *K-Nearest Neighbors (KNN)*

KNN is an instance-based non-parametric algorithm that classifies new data points based on the majority class of their k nearest neighbors in the feature space [5], [18]. It was chosen for its simplicity and ability to capture complex local decision boundaries without making assumptions about the data distribution [13].

3) *Random Forest*

Random Forest is an ensemble method that improves the weaknesses of Decision Trees through bagging (Bootstrap Aggregating) to reduce variance [19]. By building multiple decision trees on different data samples and combining their results, Random Forest achieves high accuracy and is robust against overfitting [31], [34].

4) XGBoost

XGBoost is a highly optimized and efficient implementation of gradient boosting, where trees are built sequentially and each new tree corrects the errors of the previous one [27], [8]. Features such as regularization, missing value handling, and parallel processing make it a popular and high-performance model for tabular data [34].

5) Naive Bayes (NB)

Naive Bayes (NB) is a probabilistic classification algorithm based on Bayes' Theorem, with a "naive" assumption of feature independence [35]. It calculates the probability that a document belongs to a particular sentiment class based on the words it contains [21]. Although this independence assumption simplifies computation, it can be a limitation in natural language tasks, where relationships between words are important [28].

6) Support Vector Machine (SVM)

Support Vector Machine (SVM) works by identifying the optimal hyperplane that separates data into distinct classes in a high-dimensional feature space [2], [9]. The best hyperplane is the one with the largest margin between the closest data points of each class (support vectors). The basic equation for a linear SVM classifier is:

$$w.x+b = 0$$

where w is the weight vector that defines the hyperplane's orientation, x is the feature vector, and b is the bias that adjusts the hyperplane's position [24]. To handle non-linearly separable data, SVM employs a kernel function that transforms the data into a higher-dimensional space for better separation [40].

E. Model Evaluation Metrics

After completing the modeling process, the next stage is evaluating the performance of each algorithm using the test data. Model evaluation is essential to determine how well a classifier performs in predicting unseen data and to identify potential biases or weaknesses in the model [15], [20], [34].

In this study, the evaluation metrics used are Accuracy, Precision, Recall, and F1-Score. These metrics are derived from the values in the Confusion Matrix, which consists of several components such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The mathematical formulas for each metric are presented as follows:

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

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

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

$$F1 - Score = 2 \times \frac{Precision \times Recal}{Precision + Recall}$$

Accuracy measures the overall correctness of the model's predictions, while Precision represents the proportion of correctly predicted positive samples among all predicted positives. Recall indicates the model's ability to correctly identify all positive instances in the dataset, and F1-Score provides a harmonic mean between Precision and Recall, especially useful when dealing with imbalanced datasets [15], [20], [35]. By evaluating these metrics, the study ensures a comprehensive comparison of all classification algorithms and identifies the most robust and consistent model for text-based prediction tasks [34].

Considering the imbalanced distribution of sentiment classes in the dataset (approximately 80% negative and 20% positive), this study applied a *class weighting* approach during model training, particularly for margin-based algorithms (SVM) and ensemble methods (Random Forest). This weighting technique ensures that minority classes contribute proportionally to the learning process without altering the original data distribution. Moreover, the use of the F1-Score as the primary evaluation metric further emphasizes a balanced performance assessment between precision and recall, which is particularly relevant under imbalanced data conditions.

III. RESULT AND DISCUSSIONS

This section presents the empirical findings of the study, obtained from model execution on test data. The analysis focuses on quantitative evaluation of model performance and dataset characteristics.

A. Dataset Overview

After the collection and labeling process, the final dataset consisted of 2,351 tweets related to the TPKS Law. The sentiment class distribution presented in Table 1 shows a substantial class imbalance, with negative sentiment dominating (1,886 tweets or 80.22%) compared to positive sentiment (465 tweets or 19.78%). This imbalance highlights the importance of evaluating models using not only accuracy but also complementary metrics such as Precision, Recall, and F1-Score to ensure a fair and comprehensive assessment of model performance under imbalanced conditions.

TABEL I
DISTRIBUTION OF DATASET SENTIMENT

Sentiment Class	Number of Tweets	Percentage (%)
Negative	1886	80.22
Positive	465	19.78
Total	2351	100.00

To provide a clearer overview of the dataset's textual characteristics, a word cloud visualization was generated for both positive and negative sentiment categories, as presented in Figure 2. This visualization illustrates the most frequent words extracted using the TF-IDF method, allowing readers to observe dominant lexical patterns that commonly appear in public discussions about the TPKS Law. Words such as *"korban," "kekerasan,"* and *"hukum"* frequently appear in negative sentiments, reflecting public concern about legal protection and justice for victims. Meanwhile, positive sentiments often highlight terms like *"perlindungan," "dukungan,"* and *"pemerintah,"* indicating expressions of support toward the law's implementation. This qualitative insight complements the quantitative findings by providing a contextual understanding of how public discourse linguistically manifests across sentiment categories.

Figure 2. Word Cloud of Positive and Negative Sentiments

Visualization of the most frequent words extracted using TF-IDF for positive and negative tweets related to the TPKS

Law. The word cloud highlights dominant lexical patterns in each sentiment category.

B. Exploratory Data Analysis and Feature Characteristics

An exploratory data analysis was conducted to examine the structural characteristics of the dataset and to validate the claim that the TF-IDF representation produces a high-dimensional feature space. The analysis revealed that the total number of TF-IDF features reached approximately 18,230, indicating that the text representation generated a large and sparse vector space typical of social media text data.

The average tweet length was approximately 19 words, suggesting that most tweets were short but linguistically diverse, containing a wide range of unique terms and informal expressions. A feature density histogram further confirmed that the majority of TF-IDF values were below 0.1, emphasizing the sparsity of features across documents.

To visualize the overall data structure, a Principal Component Analysis (PCA) was performed to reduce the dimensionality into two components. The 2D PCA plot revealed that the negative sentiment cluster appeared denser and more compact compared to the positive cluster, indicating that negative tweets shared more common lexical and contextual patterns. This distribution pattern supports the subsequent classification analysis by illustrating that sentiment classes are partially separable in lower-dimensional space.

Overall, this exploratory analysis highlights the complexity and high-dimensional characteristics of the dataset, providing a stronger empirical foundation for the modeling and evaluation stages presented in the following sections.

C. Model Performance Comparison

A comparative evaluation was conducted on six classification models: Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), XGBoost, K-Nearest Neighbors (KNN), and Logistic Regression (LR). All models utilized the TF-IDF method for feature extraction. The quantitative results summarized in Table 2 and visualized in Figure 3 show clear performance differences among the models.

TABEL II
COMPARISON OF CLASSIFICATION MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1-Score	AUC
LR	0.82	0.81	0.82	0.77	0.88
KNN	0.80	0.84	0.80	0.78	0.86
RF	0.85	0.85	0.85	0.82	0.91
XGBoost	0.82	0.81	0.82	0.81	0.89
NB	0.81	0.79	0.81	0.79	0.84
SVM	0.85	0.84	0.85	0.83	0.92

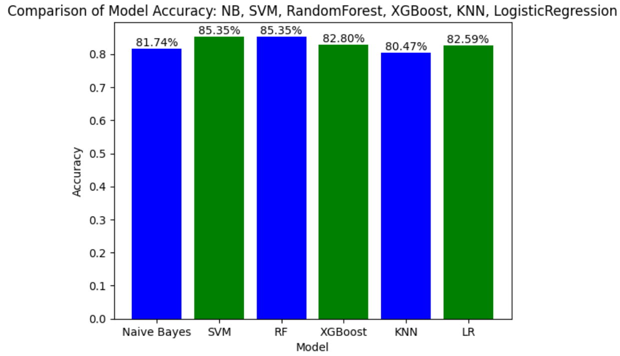

Figure 3. Comparison of Model Accuracy

The evaluation results indicate that SVM and Random Forest achieved the highest accuracy of 85.35%, outperforming the other models. These findings align with previous studies [2], [3], [9], [31], which highlight both algorithms' strong generalization capabilities in text classification tasks. SVM is effective in handling high-dimensional TF-IDF data through optimal class separation [2], [20], while Random Forest reduces overfitting by aggregating multiple decision trees [19], [31].

Meanwhile, XGBoost and Logistic Regression achieved competitive accuracy levels between 82% and 83%, consistent with research findings in [15], [24], [27], and [34]. Naive Bayes and KNN, although producing stable results, demonstrated relatively lower performance due to their sensitivity to feature independence and distance metrics [5], [18], [35].

To further strengthen the performance evaluation, Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values were analyzed for each model. Figure 4 presents the ROC curves of the two best-performing models—SVM and Random Forest.

Both models demonstrated excellent discriminative power, with AUC values of 0.92 for SVM and 0.91 for Random Forest, indicating high sensitivity and specificity in distinguishing between positive and negative sentiments. The ROC curves show that both classifiers maintain a low false positive rate while achieving a high true positive rate, confirming their robustness even under class imbalance conditions.

These findings further validate that the combination of TF-IDF features with SVM and Random Forest yields strong generalization and classification performance in Indonesian-language social media text.

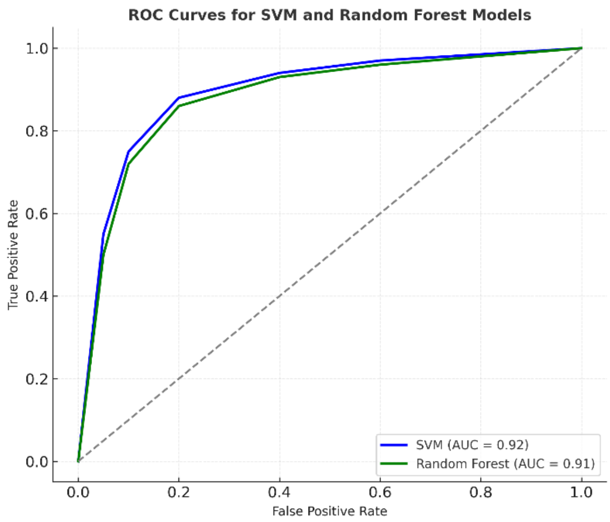

Figure 4. ROC Curves for SVM and Random Forest Models

D. Model Performance Analysis

The confusion matrix visualization (Figure 5) provides a deeper understanding of each model's predictive capability. SVM and Random Forest achieved the highest true positive and true negative values, confirming their robustness in identifying sentiment polarity. Specifically, SVM recorded 369 correct predictions for the negative class and 33 for the positive class, while Random Forest achieved 374 correct predictions for the negative class and 29 for the positive class.

In contrast, Naive Bayes and KNN exhibited higher false-negative counts, suggesting limitations in capturing the nuanced contextual features of social media text.

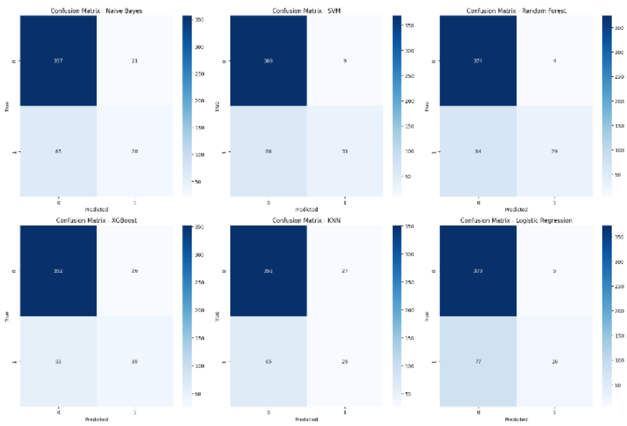

Figure 5. Confusion Matrix of Classification Models

These results demonstrate that the TF-IDF feature representation was sufficiently informative to separate sentiment classes, especially when combined with robust algorithms such as SVM and Random Forest. Nevertheless, models like Naive Bayes and KNN were less effective because the assumptions of feature independence and distance similarity do not fully apply to informal Indonesian text data.

SVM is able to maximize the inter-class margin in the high-dimensional feature space generated by TF-IDF, making it effective for text with many sparse features. Meanwhile, Random Forest excels due to its bagging and feature randomness mechanisms, which reduce variance and increase stability in data with large dimensions and imbalanced class distributions. Furthermore, an interpretation of the confusion matrix is provided to demonstrate how both models minimize Type I and II errors.

To statistically validate whether the observed performance differences among the classification algorithms are meaningful rather than random variations, a paired t-test was conducted on the F1-Score results obtained from the 10-Fold Cross-Validation. The statistical test revealed that both Support Vector Machine (SVM) and Random Forest (RF) significantly outperformed the other algorithms ($p < 0.05$), confirming the robustness of their predictive capabilities. However, the difference between SVM and Random Forest was not statistically significant ($p = 0.27$), indicating that their performance levels are relatively comparable. These results reinforce the conclusion that both SVM and Random Forest exhibit superior and consistent generalization ability for Indonesian-language sentiment analysis tasks, particularly in handling high-dimensional and imbalanced datasets.

E. Feature Extraction Discussion

This study employed the Term Frequency–Inverse Document Frequency (TF-IDF) method as the primary feature extraction approach.

In this study, the TF-IDF feature extraction was implemented with specific parameter settings to ensure an optimal balance between feature diversity and computational efficiency. The configuration used was max_features = 5000 to limit the vocabulary size to the most informative words, ngram_range = (1, 2) to include both unigrams and bigrams for capturing short contextual patterns, and min_df = 3 and max_df = 0.8 to remove extremely rare or overly frequent terms. These parameter values were selected empirically to enhance the representational quality of the text while preventing overfitting and maintaining interpretability.

TF-IDF effectively converted unstructured tweet texts into numeric vectors that emphasize the most informative terms, allowing the models to distinguish between positive and negative sentiments. These results are consistent with previous studies [7], [17], [34], which confirmed that TF-IDF provides a good balance between computational simplicity and classification accuracy.

However, TF-IDF has inherent limitations in capturing semantic relationships between words [9], [40]. Future research can address this issue by implementing N-gram modeling to capture phrase-level sentiment patterns, applying feature selection methods such as Chi-Square or Mutual Information to eliminate irrelevant terms [19], or utilizing word embedding techniques such as FastText or IndoBERT for richer contextual representation [22], [40].

In this study, TF-IDF remains a suitable choice because it provides an effective balance between performance and computational efficiency, particularly for large-scale social media sentiment analysis.

F. Overall Discussion

The overall findings confirm that Support Vector Machine (SVM) and Random Forest (RF) are the most effective models for analyzing public sentiment toward the TPKS Law on Twitter. Both models achieved high accuracy, precision, recall, and F1-score values. Their robustness can be attributed to their ability to manage high-dimensional, imbalanced, and noisy textual data effectively.

The superior performance of SVM and Random Forest can be theoretically explained by the intrinsic characteristics of the TF-IDF feature space and the fundamental learning principles of both algorithms. TF-IDF produces a high-dimensional and sparse representation that tends to exhibit partial linear separability, a condition that suits the margin-based optimization in SVM. By maximizing the separating hyperplane, SVM achieves strong generalization even when data sparsity is high. Conversely, Random Forest effectively handles redundant and noisy TF-IDF features through ensemble averaging. Its random feature selection and bootstrap aggregation mechanisms mitigate variance and overfitting, leading to stable results in heterogeneous text datasets. These theoretical insights clarify why both models consistently outperform other algorithms in this study.

The combination of TF-IDF with SVM or Random Forest has proven to be a reliable and efficient approach for sentiment classification in Indonesian-language datasets. Furthermore, these findings reaffirm the role of machine learning as a powerful analytical tool for understanding public perception of government policies using data-driven methods [7], [17], [21].

From a theoretical perspective, the superior performance of SVM can be explained through the concept of *linear separability* inherent in TF-IDF vector spaces. Since TF-IDF transforms text into high-dimensional sparse representations, SVM effectively identifies the optimal hyperplane that separates sentiment classes with maximum margin, thereby reducing classification errors. This observation aligns with Sarker [16], who emphasizes that SVM is particularly suitable for text classification tasks involving sparse and high-dimensional data, where the linear kernel can maximize generalization capability.

In contrast, the strong performance of Random Forest can be interpreted through the *bias–variance trade-off* principle. By aggregating multiple decision trees generated from bootstrapped samples, Random Forest reduces model variance while maintaining low bias, resulting in more stable predictions across diverse textual inputs. This explanation is consistent with Adugna et al. [31], who found that ensemble-based approaches tend to outperform individual classifiers when handling noisy, heterogeneous data such as social media text.

Finally, the inclusion of six classification algorithms in this study reflects the essence of the *No Free Lunch Theorem*, which asserts that no single algorithm performs best across all problem domains. Consequently, comparing multiple algorithms provides methodological rigor and empirical validation in identifying models that best fit the characteristics of Indonesian-language sentiment data on Twitter.

G. Computational Efficiency

In addition to accuracy, computational efficiency is an important aspect for assessing the practical feasibility of each model. The training time for all algorithms was measured using a personal computer with the following specifications: Processor: Intel(R) Celeron(R) CPU 4305U @ 2.20GHz (2.21 GHz); Installed RAM: 4.00 GB (3.86 GB usable); System Type: 64-bit operating system, x64-based processor. The training time results are summarized in Table 3.

TABEL III
TRAINING TIME AND MODEL EFFICIENCY

Model	**Training Time (seconds)**	**Description**
Logistic Regression	10	Fast
KNN	25	Relatively slow (affected by data size)
Random Forest	42	Fairly high (due to multiple decision trees)
XGBoost	48	Highest (iterative and complex process)
Naive Bayes	7	Very fast
SVM	31	Moderate

The results show that Naive Bayes and Logistic Regression are the fastest models, making them suitable for real-time applications or systems with limited resources. Meanwhile, SVM and Random Forest require longer training times but provide superior predictive performance, indicating a trade-off between accuracy and computational cost. These findings emphasize that algorithm selection should balance accuracy, model complexity, and available computing capacity according to the intended application context.

H. Research Limitations

This study has several limitations that should be acknowledged as a foundation for future research development. The dataset used in this study is relatively small, consisting of 2,351 tweets collected only during the 2022–2025 period. This limited sample size may restrict the generalizability of the findings to a broader range of social and temporal contexts. Moreover, the labeling process was conducted manually, which may introduce subjective bias due to differences in annotator interpretation. These factors could influence the consistency and objectivity of sentiment classification, particularly in complex social discourse such as the TPKS Law discussion.

Furthermore, the use of the TF-IDF representation method has limitations in capturing the semantic context and relationships between words, which may constrain the model's ability to interpret implicit meanings in informal language expressions on social media. In addition, the models developed in this study have not yet been evaluated using real-time or streaming data, leaving their performance in dynamic or continuously updated environments uncertain. Future research is expected to overcome these limitations by expanding dataset coverage, applying automated or crowdsourced labeling strategies, adopting embedding-based representations such as FastText or IndoBERT to enhance semantic understanding, and implementing real-time sentiment monitoring systems for continuous public opinion analysis.

IV. CONCLUSION

This study successfully achieved its objective of identifying the most effective machine learning algorithms for sentiment analysis of public opinion toward the TPKS Law on Twitter. A comparative evaluation of six algorithms demonstrated that Support Vector Machine (SVM) and Random Forest (RF) achieved the highest accuracy (85.35%) and F1-score (0.83), confirming their superiority in handling high-dimensional and imbalanced textual data. Statistical testing using a paired t-test ($p < 0.05$) validated that both models significantly outperformed other algorithms. Furthermore, efficiency analysis revealed that SVM and RF offer an optimal balance between computational cost and predictive accuracy. The overall sentiment distribution showed that public perception of the TPKS Law remains predominantly negative (80.22%), highlighting the importance of continuous public communication regarding policy implementation. Methodologically, the findings reaffirm the effectiveness of combining TF-IDF with margin-based and ensemble classifiers for Indonesian-language text analytics. Future research may enhance this approach by employing transformer-based architectures such as IndoBERT or IndoBERTweet to capture deeper semantic nuances, and by applying real-time sentiment monitoring for policy evaluation. These contributions provide both theoretical and practical value to the field of computational social science and public policy informatics in Indonesia.

REFERENCES

[1] A. Jannat, U. H. Abu, N. Tahreem, and M. Tayyaba, "A Review On Software Testing And Its Methodology," *i-manager's J. Softw. Eng.*, vol. 13, no. 3, p. 32, 2019, doi: 10.26634/jse.13.3.15515.

[2] E. A. Shams and A. Rizaner, "A novel support vector machine based intrusion detection system for mobile ad hoc networks," *Wirel. Networks*, vol. 24, no. 5, pp. 1821–1829, Jul. 2018, doi: 10.1007/s11276-016-1439-0.

[3] S. Aljawarneh, M. Aldwairi, and M. B. Yassein, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model," *J. Comput. Sci.*, vol. 25, pp. 152–160, Mar. 2018, doi: 10.1016/j.jocs.2017.03.006.

[4] Y. I. Kurniawan, A. Rahmawati, N. Chasanah, and A. Hanifa, "Application for determining the modality preference of student learning," *J. Phys. Conf. Ser.*, vol. 1367, no. 1, p. 012011, Nov. 2019, doi: 10.1088/1742-6596/1367/1/012011.

[5] Y. Guo, S. Han, Y. Li, C. Zhang, and Y. Bai, "K-Nearest Neighbor combined with guided filter for hyperspectral image classification," *Procedia Comput. Sci.*, vol. 129, pp. 159–165, 2018, doi: 10.1016/j.procs.2018.03.066.

[6] Y. I. Kurniawan, E. Soviana, and I. Yuliana, "Merging Pearson Correlation and TAN-ELR algorithm in recommender system," 2018, p. 040028. doi: 10.1063/1.5042998.

[7] E. A. Sukma, A. N. Hidayanto, A. I. Pandesenda, A. N. Yahya, P. Widharto, and U. Rahardja, "Sentiment Analysis of the New Indonesian Government Policy (Omnibus Law) on Social Media Twitter," in *2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, IEEE, Nov. 2020, pp. 153–158. doi: 10.1109/ICIMCIS51567.2020.9354287.

[8] M. F. Almufareh, N. Jhanjhi, M. Humayun, G. N. Alwakid, D. Javed, and S. N. Almuayqil, "Integrating Sentiment Analysis with Machine Learning for Cyberbullying Detection on Social Media," *IEEE Access*, vol. 13, no. May, pp. 78348–78359, 2025, doi: 10.1109/ACCESS.2025.3558843.

[9] Z. Al Faridzi, D. Pramesti, and R. Y. Fa'rifah, "A Comparison of Oversampling and Undersampling Methods in Sentiment Analysis Regarding Indonesia Fuel Price Increase Using Support Vector Machine," in *2023 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS)*, IEEE, Aug. 2023, pp. 1–6. doi: 10.1109/ICADEIS58666.2023.10270851.

[10] N. Agustina, A. Hadhiwibowo, I. K. A. S. Saputra, W. Cahyati, Z. Pupi Indriarti, and B. Rustandi, "Sentiment Analysis of Air Quality to Identify Public Perspectives in Indonesia Based on Machine Learning," in *2023 17th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/TSSA59948.2023.10366899.

[11] Gilbert, S. Beatricia, A. Reinaldo, and M. F. Hasani, "Text Augmentation for Indonesian Intent Classification: Comparative Study," in *2024 11th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, IEEE, Aug. 2024, pp. 265–270. doi: 10.1109/ICITACEE62763.2024.10762810.

[12] G. E. Agustine and R. Jayadi, "Sentiment Analysis of Bamboo Charcoal: Comparing Machine Learning Algorithms for Effective Insights," *J. La Multiapp*, vol. 6, no. 2, pp. 287–298, Mar. 2025, doi: 10.37899/journallamultiapp.v6i2.2010.

[13] M. Febima and L. Magdalena, "Predictive Analytics on Shopee for Optimizing Product Demand Prediction through K-Means Clustering and KNN Algorithm Fusion," *J. Inf. Syst. Informatics*, vol. 6, no. 2, pp. 751–765, Jun. 2024, doi: 10.51519/journalisi.v6i2.720.

[14] L. Medeiros Neto, S. Rogerio da Silva Neto, and P. T. Endo, "A comparative analysis of converters of tabular data into image for the classification of Arboviruses using Convolutional Neural Networks," *PLoS One*, vol. 18, no. 12, p. e0295598, Dec. 2023, doi: 10.1371/journal.pone.0295598.

[15] B. B. Acharya, "Comparative Analysis of Machine Learning Algorithms: KNN, SVM, Decision Tree and Logistic Regression for Efficiency and Performance," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. 11, pp. 614–619, Nov. 2024, doi: 10.22214/ijraset.2024.65138.

[16] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.

[17] R. Nurdiyanti and E. Utami, "Benchmarking Multiple Machine Learning Algorithms for Sentiment Analysis on Sexual Violence," in *2024 6th International Conference on Cybernetics and Intelligent System (ICORIS)*, IEEE, Nov. 2024, pp. 1–6. doi:

10.1109/ICORIS63540.2024.10903985.

[18] R. Wahyusari, S. Sunardi, and A. Fadlil, “Comparative Analysis Of Ant Lion Optimization And Jaya Algorithm For Feature Selection In K-Nearest Neighbor (Knn) Based Electricity Consumption Prediction,” *J. Tek. Inform.*, vol. 6, no. 3, pp. 1373–1388, Jun. 2025, doi: 10.52436/1.jutif.2025.6.3.4692.

[19] X. Yuan, S. Liu, W. Feng, and G. Dauphin, “Feature Importance Ranking of Random Forest-Based End-to-End Learning Algorithm,” *Remote Sens.*, vol. 15, no. 21, p. 5203, Nov. 2023, doi: 10.3390/rs15215203.

[20] E. M. A. Allah, D. E. El-Matary, E. M. Eid, and A. S. T. El Dien, “Performance Comparison of Various Machine Learning Approaches to Identify the Best One in Predicting Heart Disease,” *J. Comput. Commun.*, vol. 10, no. 02, pp. 1–18, 2022, doi: 10.4236/jcc.2022.102001.

[21] Y. Ansori and K. F. H. Holle, “Perbandingan Metode Machine Learning dalam Analisis Sentimen Twitter,” *J. Sist. dan Teknol. Inf.*, vol. 10, no. 4, p. 429, Dec. 2022, doi: 10.26418/justin.v10i4.51784.

[22] K. K. Agustiningsih, E. Utami, and M. A. Alsyaibani, “Sentiment Analysis of COVID-19 Vaccines in Indonesia on Twitter Using Pre-Trained and Self-Training Word Embeddings,” *J. Ilmu Komput. dan Inf.*, vol. 15, no. 1, pp. 39–46, Feb. 2022, doi: 10.21609/jiki.v15i1.1044.

[23] S. Aslan, S. Kızıloluk, and E. Sert, “TSA-CNN-AOA: Twitter sentiment analysis using CNN optimized via arithmetic optimization algorithm,” *Neural Comput. Appl.*, vol. 35, no. 14, pp. 10311–10328, May 2023, doi: 10.1007/s00521-023-08236-2.

[24] R. R. Waliyansyah and Umar Hafidz Asy’ari Hasbullah, “Comparison of Tree Method, Support Vector Machine, Naïve Bayes, and Logistic Regression on Coffee Bean Image,” *Emit. Int. J. Eng. Technol.*, vol. 9, no. 1, pp. 126–136, Jun. 2021, doi: 10.24003/emitter.v9i1.536.

[25] O. Abiola, A. Abayomi-Alli, O. A. Tale, S. Misra, and O. Abayomi-Alli, “Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser,” *J. Electr. Syst. Inf. Technol.*, vol. 10, no. 1, p. 5, Jan. 2023, doi: 10.1186/s43067-023-00070-9.

[26] A. L. Qosim, F. Kurniawan, U. Bahruddin, Z. Mubaraq, Suhartono, and M. Faisal, “Analysis Classification Opinion of Policy Government Announces Cabinet Reshuffle on YouTube Comments Using 1D Convolutional Neural Networks,” in *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, IEEE, Apr. 2021, pp. 30–35. doi: 10.1109/EIConCIT50028.2021.9431884.

[27] M. R. Kurniawanda and F. A. T. Tobing, “Analysis Sentiment Cyberbullying In Instagram Comments with XGBoost Method,” *IJNMT (International J. New Media Technol.*, vol. 9, no. 1, pp. 28–34, Jul. 2022, doi: 10.31937/ijnmt.v9i1.2670.

[28] H. Fouadi, H. El Moubtahij, H. Lamtougui, and A. Yahyaouy, “Sentiment Analysis Of Arabic Comments Using Machine Learning And Deep Learning Models,” *Indian J. Comput. Sci. Eng.*, vol. 13, no. 3, pp. 598–606, Jun. 2022, doi: 10.21817/indjcse/2022/v13i3/221303003.

[29] Aiswarya A S and H. Rajeev, “Youtube Comment Sentimental Analysis,” *Indian J. Data Min.*, vol. 4, no. 1, pp. 5–8, May 2024, doi: 10.54105/ijdm.A1633.04010524.

[30] S. N. Alsubari, “Data Analytics for the Identification of Fake Reviews Using Supervised Learning,” *Comput. Mater. Contin.*, vol. 70, no. 2, pp. 3189–3204, 2022, doi: 10.32604/cmc.2022.019625.

[31] T. Adugna, W. Xu, and J. Fan, “Comparison of Random Forest and Support Vector Machine Classifiers for Regional Land Cover Mapping Using Coarse Resolution FY-3C Images,” *Remote Sens.*, vol. 14, no. 3, p. 574, Jan. 2022, doi: 10.3390/rs14030574.

[32] D. K. Dake and E. Gyimah, “Using sentiment analysis to evaluate qualitative students’ responses,” *Educ. Inf. Technol.*, vol. 28, no. 4, pp. 4629–4647, Apr. 2023, doi: 10.1007/s10639-022-11349-1.

[33] A. Taner *et al.*, “Multiclass Apple Varieties Classification Using Machine Learning with Histogram of Oriented Gradient and Color Moments,” *Appl. Sci.*, vol. 13, no. 13, p. 7682, Jun. 2023, doi: 10.3390/app13137682.

[34] F. U. Aulya and K. Kusnawi, “Evaluating Classification Models for Predicting Product Success in Indonesian E-Commerce,” *J. Tek. Inform.*, vol. 6, no. 4, pp. 2723–2739, Aug. 2025, doi: 10.52436/1.jutif.2025.6.4.5071.

[35] Dewi Widyawati and Amaliah Faradibah, “Comparison Analysis of Classification Model Performance in Lung Cancer Prediction Using Decision Tree, Naive Bayes, and Support Vector Machine,” *Indones. J. Data Sci.*, vol. 4, no. 2, pp. 80–89, Jul. 2023, doi: 10.56705/ijodas.v4i2.76.

[36] J. Xue *et al.*, “Using Twitter-Based Data for Sexual Violence Research: Scoping Review,” *J. Med. Internet Res.*, vol. 25, p. e46084, May 2023, doi: 10.2196/46084.

[37] A. S. Khan, A. Bibi, A. Khan, and I. Ahmad, “Responsibility of Sexual Violence Under International Law,” *J. Soc. Sci. Rev.*, vol. 3, no. 1, pp. 29–41, Mar. 2023, doi: 10.54183/jssr.v3i1.110.

[38] N. Akram and M. Yasmin, “Sexual violence against women: Global interventions and an evidence from Pakistan,” *Womens. Stud. Int. Forum*, vol. 97, p. 102691, Mar. 2023, doi: 10.1016/j.wsif.2023.102691.

[39] T. H. Rochadiani, “Sentiment Analysis of YouTube Comments Toward Chat GPT,” *J. Transform.*, vol. 21, no. 1, Aug. 2023, doi: 10.26623/transformatika.v21i2.7033.

[40] B. AlBadani, R. Shi, and J. Dong, “A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM,” *Appl. Syst. Innov.*, vol. 5, no. 1, p. 13, Jan. 2022, doi: 10.3390/asi5010013.

[41] B. R. Chakravarthi *et al.*, “DravidianCodeMix: sentiment analysis and offensive language identification dataset for Dravidian languages in code-mixed text,” *Lang. Resour. Eval.*, vol. 56, no. 3, pp. 765–806, Sep. 2022, doi: 10.1007/s10579-022-09583-7.

[42] L. Holla and K. S. Kavitha, “An Improved Fake News Detection Model Using Hybrid Time Frequency-Inverse Document Frequency for Feature Extraction and AdaBoost Ensemble Model as a Classifier,” *J. Adv. Inf. Technol.*, vol. 15, no. 2, pp. 202–211, 2024, doi: 10.12720/jait.15.2.202-211.