Informatics, Electrical and Electronics Engineering (Infotron)
Volume 4, Nomor 2, (2024) pp. 63-78

ISSN 2798-0197
http://riset.unisma.ac.id/index.php/infotron/article/view/XXXXX

Comparative Analysis of Machine Learning Models for Identifying Cybercrimes in Social Media Comments

Abd. Charis Fauzan[a], Mochammad Arifin[b], Veradella Yuelisa Mafula[c]

[a,c]*Departments of Computer Science, Universitas Nahdlatul Ulama Blitar, Kota Blitar, Indonesia*
[b]*Departments of Islamic Family Law, Universitas Nahdlatul Ulama Blitar, Kota Blitar, Indonesia*

email: [a]abdcharis@unublitar.ac.id, [b]mochammadarifin@unublitar.ac.id, [c]veradella@unublitar.ac.id*

*Corresponding Author

ARTICLE INFORMATION

Article history:
Received 31 October 2024
Reviewed 5 November 2024
Revised 25 November 2024
Accepted 30 November 2024

Keywords:
Comparative Analysis,
Cybercrimes,
Machine Leaning,
Social Media Comments

A B S T R A CT

The rapid growth of social media has created opportunities for digital interaction but has also introduced challenges, particularly in addressing cybercrimes such as defamation, threats, and SARA-related content. Cybercrime detection on social media is critical as it helps mitigate the spread of harmful behavior, safeguard users, and support law enforcement in addressing violations like Indonesia's Information and Electronic Transactions Law (UU ITE). This study conducts a comparative analysis of machine learning algorithms—Naive Bayes, Support Vector Machines (SVM), and Random Forests—to identify cybercrimes in social media comments. Using a sentiment-labeled dataset obtained from Kaggle, consisting of Indonesian social media comments from Twitter (X), the comments are categorized into seven specific classes: Neutral Sentiment, Positive Sentiment, Negative Sentiment, Insulting Government, Insulting or Defaming Others, Threatening Others, and SARA-Based Content. The results show that Random Forest achieved the highest overall accuracy (91%) and performed best in detecting moderately represented classes such as Insulting Government. SVM demonstrated robust performance with 88% accuracy, particularly excelling in identifying dominant classes like Negative Sentiment, while Naive Bayes, though computationally efficient, struggled with minority classes, achieving an accuracy of 73%. However, the dataset's imbalance posed challenges for all algorithms, particularly with underrepresented categories. This limitation underscores the need for more diverse and representative datasets to improve model performance and ensure broader applicability of the findings.

1. Introduction

The rapid development of information and communication technology has significantly transformed the way people interact and communicate globally [1]. Social media platforms, as one of the most dominant tools for communication and information sharing, have facilitated unprecedented opportunities for connectivity and interaction [2]. However, the widespread adoption of social media has also introduced challenges, particularly in addressing cybercrimes such as hate speech, defamation, and cyberbullying. These harmful behaviors threaten social harmony and the integrity of online discourse, posing substantial challenges for governments and legal institutions tasked with maintaining societal order and enforcing legal frameworks [3]. In Indonesia, the situation is particularly concerning, given the increasing number of cases involving violations of the Information and Electronic Transactions Law (UU ITE), which regulates online behavior [4]. Social media users often exploit the anonymity and reach of these platforms to engage in unlawful activities, including spreading hate speech, defaming public figures, and propagating misinformation [5]. For instance, reports indicate that online hate speech cases have risen sharply, with significant portions of this activity targeting ethnic and religious groups. The massive scale of content generated daily on these platforms renders manual

moderation methods insufficient, necessitating automated and scalable solutions to ensure compliance with legal norms and promote ethical digital interactions [6].

The urgency of addressing these issues stems from the profound impact that cybercrimes have on social cohesion and the rule of law. Hate speech and defamatory remarks can polarize communities, erode trust in public institutions, and incite violence. Meanwhile, cyberbullying can lead to psychological harm and social exclusion, particularly among vulnerable populations such as children and young adults [7]. Left unchecked, these activities not only harm individuals but also undermine societal stability, making it imperative to develop systems capable of detecting and categorizing legal violations in social media comments. Such systems are critical for fostering a safe and ethical digital environment [8]. This study explores the use of Artificial Intelligence (AI), specifically machine learning algorithms, to detect and classify cybercrimes in social media comments. Machine learning offers a powerful and efficient alternative to manual moderation, enabling the processing of large-scale social media data in real-time. By automating the identification of harmful content, machine learning can support governments and social media platforms in enforcing digital laws and promoting responsible online interactions [9].

Machine learning techniques have shown great promise in text classification and sentiment analysis tasks [10]. For instance, Naive Bayes, Support Vector Machines (SVM), and Random Forests have been widely used to identify hate speech and toxic comments with high accuracy. Arfan, *et al*, for example, successfully employed Naive Bayes to detect offensive language in various datasets, achieving notable performance across different contexts [11]. Similarly, SVM has demonstrated robust performance in multilingual datasets, effectively distinguishing harmful content from benign interactions [12]. Random Forests, known for their versatility, have been applied in hybrid models combining linguistic features with machine learning techniques, improving precision in detecting cyberbullying and other harmful behaviors [13]. Despite their efficacy, many studies focus solely on general text classification without adapting their approaches to specific legal frameworks. This limitation is particularly evident in the context of Indonesia, where the UU ITE outlines explicit categories of unlawful online behavior, including hate speech, defamation, and threats [14]. By failing to tailor their methods to these legal requirements, existing studies miss the opportunity to provide practical tools for law enforcement and digital governance. Moreover, while previous research has highlighted the effectiveness of AI in detecting toxic content, few have explored its integration with specific legal frameworks to enhance its practical applicability [15].

This research aims to fill this gap by integrating machine learning techniques with a legal perspective to identify cybercrimes in Indonesian social media comments. A sentiment-labeled dataset, specifically tailored to the categories defined by the UU ITE, serves as the foundation for this study. Naive Bayes, SVM, and Random Forests are employed to detect harmful behaviors and classify legal violations. These algorithms were chosen based on their established efficacy in handling text classification tasks and their contrasting strengths [16]. Naive Bayes is known for computational efficiency and suitability for real-time applications, SVM for its robustness in managing high-dimensional data, and Random Forests for balancing accuracy and interpretability. Comparing these methods provides a comprehensive understanding of their performance across diverse cybercrime categories, addressing the limitations of prior studies that focused on single algorithms without exploring their relative strengths and weaknesses. The urgency of this study lies in its ability to bridge the gap between machine learning advancements and practical legal enforcement. By enabling real-time monitoring and filtering of harmful content, the proposed system supports the enforcement of Indonesia's UU ITE, fostering a safer and more ethical digital environment. This comparative approach underscores the critical role of AI in promoting responsible online behavior and advancing digital citizenship principles.

2. Method

Based on Figure 1, this research begins with dataset preparation, where social media comments are collected and labeled based on 7 categories such as neutral sentiment, positive sentiment, negative sentiment, insulting government, threatening others, and SARA-based content. This dataset provides the foundation for subsequent analyses. Next, text preprocessing is conducted to clean and prepare the

raw text data. This involves removing noise such as URLs, punctuation, and stopwords while converting the text to lowercase. Additionally, tokenization is performed to break the text into manageable components for machine learning processing. To address the challenge of class imbalance, handling class imbalance techniques like SMOTE (Synthetic Minority Oversampling Technique) are applied, ensuring that minority classes are sufficiently represented during training. This step is crucial to improve model performance across all categories. The model training and evaluation phase involves applying the three algorithms using Naive Bayes, Support Vector Machine and Random Forest. Each model's performance is evaluated based on accuracy, precision, recall, and F1-score. Finally, comparative analysis highlights the strengths and weaknesses of each algorithm, identifying the most suitable approach for detecting cybercrimes in social media comments.

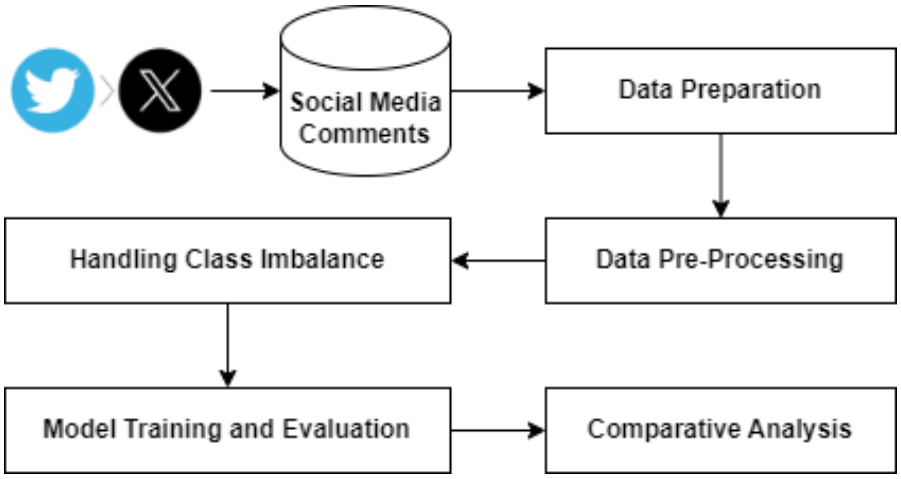

Fig 1. Research Method

2.1 Dataset Preparation

The dataset consists of social media comments collected from Twitter, specifically curated to support the identification and classification of cybercrimes as outlined under Indonesia's Information and Electronic Transactions Law (UU ITE). The dataset is sourced from Kaggle, titled Indonesia Twitter Comment Labeled with ITE Law. It has been meticulously annotated to align with specific legal articles in the Indonesian UU ITE and Criminal Code (KUHP). The dataset includes various types of comments, ranging from neutral and positive interactions to harmful or illegal content. Each comment is systematically labeled into one of seven distinct categories, reflecting both general sentiment and potential legal violations.

Table 1. Dataset Summary

Data Labels	Description	Examples	Labeled Comments
Neutral Sentiment (0)	Objective, neutral comments without any legal violations.	*"Barusan liat tulisan di belakang truk rela injak kaki demi sesuap nasi."*	5327
Positive Sentiment (1)	Comments that convey support, positivity, or encouragement	*"Doa rezeki tak putus, semoga semua bahagia!"*	2792
Negative Sentiment (2)	Comments expressing criticism or dissatisfaction, without explicit legal violations.	*"Gua dah kabor nak bagi gua specific task jangan suka modus."*	4188
Insulting Government (3)	Comments insulting public institutions, violating *Pasal 207/208 KUHP.*	*"Hei kunyuk USER, ente paham gak? Pemerintah ini cuma omdo!"*	58

Insulting or Defaming Others (4)	Personal attacks or defamation under *Pasal 27 ayat (3) UU ITE.*	*"Dia itu pembohong besar, cuma mau cari perhatian!"*	152
Threatening Others (5)	Comments that intimidate or threaten individuals, violating *Pasal 29 UU ITE.*	*"Kalau lo berani lagi ngomong, gua samperin rumah lo!"*	29
SARA-Based Content (6)	Hate speech targeting ethnicity, religion, race, or intergroup differences, under Pasal 28 ayat (2) UU ITE.	*"Agama kalian itu paling bodoh, cuma bikin keributan."*	100
All categories labels			**12647**

Based on Table 1, the dataset summary provided in the table above describes a collection of 12,647 labeled comments from Twitter, which are categorized into 7 labels based on their sentiment and alignment with specific legal definitions under Indonesia's Information and Electronic Transactions Law (UU ITE) and the Criminal Code (KUHP).

2.2 Data Pre-Processing

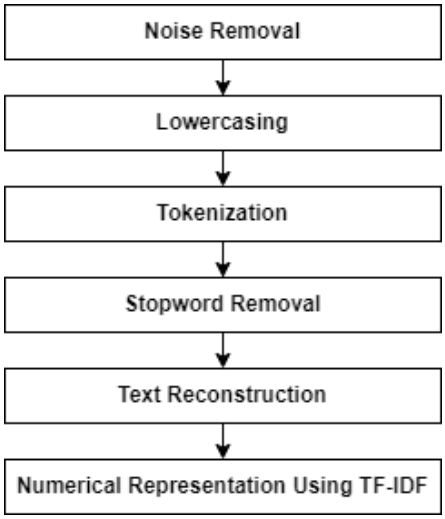

Fig 2. Preprocessing Stages

Text preprocessing is a critical step in preparing textual data for machine learning models, as raw text data from social media platforms like Twitter often contains noise, irrelevant information, and inconsistencies. In this study, the preprocessing stage is designed to clean, standardize, and transform the data into a structured format suitable for analysis and model training. The diagram in Figure 2 outlines the sequential stages of text preprocessing. The preprocessing stages include following steps:

2.2.1 Noise Removal

This initial stage involves eliminating unnecessary elements such as URLs, punctuation, and numerical characters from the raw text. The goal is to reduce noise and retain only meaningful content relevant to the analysis.

2.2.2 Lowercasing

After noise removal, all text is converted to lowercase to ensure uniformity. This avoids treating words like "Government" and "government" as distinct, improving consistency across the dataset.

2.2.3Tokenization

The text is then broken down into individual words, or tokens, which allows each word to be processed independently. For example, the phrase “I love Indonesia” becomes ['i', 'love', 'indonesia'].

2.2.4 Stopword Removal

Commonly used words such as "and," "the," or "is," which add little contextual value, are filtered out. This step focuses the analysis on more significant terms that contribute to sentiment or meaning.

2.2.5 Text Reconstruction

The remaining tokens are reassembled into a cleaned text string, ensuring coherence while maintaining the filtered structure.

2.2.6 Numerical Representation Using TF-IDF

Finally, the cleaned text is transformed into numerical vectors using Term Frequency-Inverse Document Frequency (TF-IDF). This step quantifies the importance of each word relative to its frequency across the dataset [17], enabling machine learning algorithms to process the data effectively.

2.3 Handling Class Imbalance

Handling class imbalance refers to addressing the disproportionate distribution of data across different labels in a dataset [18]. From the table, it is evident that the majority of the labeled comments belong to "Neutral Sentiment" (5,327 comments) and "Negative Sentiment" (4,188 comments), while minority classes such as "Threatening Others" (29 comments) and "Insulting Government" (58 comments) are significantly underrepresented. This imbalance can lead to biased machine learning models that favor majority classes while ignoring critical minority class samples. In this research, the minority classes—such as "Threatening Others" or "SARA-Based Content"—are critical for detecting cybercrimes in social media. If class imbalance is not addressed, the model will likely underperform in identifying these minority categories, leading to poor recall rates for important cases like threats or hate speech. This failure would render the model ineffective for real-world applications, where detecting harmful behavior is essential for enforcing legal compliance and promoting responsible online interactions. The Synthetic Minority Oversampling Technique (SMOTE) is used to balance the dataset by synthetically generating samples for minority classes [19]. The process involves:

2.3.1 Identifying Imbalance

Analyze the dataset to determine the distribution of each label. Minority classes, such as "Threatening Others" (29) and "SARA-Based Content" (100), are highlighted as underrepresented.

2.3.2 Applying SMOTE

For each minority class, synthetic samples are created by interpolating between existing data points. SMOTE generates these samples by selecting k-nearest neighbors and creating new samples between them using the Equation (1).

$$x_{\text{new}} = x_i + \lambda \times (x_{\text{neighbor}} - x_i) \tag{1}$$

Here, x_i is an existing data point, $x_{neighbor}$ is a nearby point, and λ is a random number between 0 and 1. The oversampled minority classes are combined with the majority classes, creating a balanced training dataset where all categories have comparable representation.

2.4 Training Model and Evaluation

This research focuses on using Naive Bayes, Support Vector Machines (SVM), and Random Forest algorithms to classify social media comments into six sentiment categories. These range from neutral and positive sentiments to harmful content like hate speech, threats, and defamation. The dataset, originally imbalanced (e.g., "Neutral Sentiment" has 5,327 samples, while "Threatening Others" has only 29), was balanced using the SMOTE method before training the models. This ensures fair representation of minority classes, critical for detecting cybercrimes.

2.4.1 Naive Bayes

Naive Bayes is a probabilistic algorithm based on Bayes' Theorem [20][21], which calculates the posterior probability of a x as Equation (2).

$$P(C_k|x) = \frac{P(x|C_k) \cdot P(C_k)}{P(x)} \tag{2}$$

Using TF-IDF vectors generated from the text preprocessing stage, Naive Bayes computes the probabilities of each class for a given comment. This study employs Multinomial Naive Bayes (MultinomialNB), which is specifically designed to handle data in discrete distributions, such as frequency-based text representations like TF-IDF [22]. Handling imbalance ensures the algorithm does not overly favor majority classes. While Naive Bayes is efficient and scalable, it assumes feature independence, which can limit its ability to model complex relationships in the text. This assumption, while simplifying computations, may not fully capture nuanced patterns in natural language data.

2.4.2 SVM

SVM is a supervised learning algorithm that finds the optimal hyperplane separating classes in a high-dimensional feature space. The hyperplane is defined as Equation (3).

$$\mathrm{w}^{\mathrm{T}}\mathrm{x} + \mathrm{b} = 0 \tag{3}$$

SVM maximizes the margin, the distance between the hyperplane and the nearest data points (support vectors). In this study, SVM utilizes a linear kernel to separate the data using a linear hyperplane. Linear kernels are particularly well-suited for high-dimensional data, such as text, where the number of features (words) can be very large. This approach efficiently handles the sparse and high-dimensional nature of text data, especially when represented using techniques like TF-IDF [23]. By balancing the dataset using SMOTE, the minority classes like "Threatening Others" (originally underrepresented) gain equal importance during training. This improves the model's ability to classify rare yet critical categories. SVM's ability to perform well in imbalanced datasets, combined with the linear kernel's effectiveness in handling large feature spaces, makes it highly suitable for text classification tasks.

2.4.3 Random Forest

Random Forest is an ensemble method that builds multiple decision trees on random subsets of the data and combines their outputs using majority voting [24]. The final prediction is given as Equation (4)

$$y_{\text{predicted}} = \text{Mode}(T_1(x), T_2(x), \ldots, T_n(x)) \tag{4}$$

The balanced dataset generated by SMOTE ensures that minority classes like "Insulting Government" (58 samples) are adequately represented during training. Random Forest excels at handling diverse patterns in the dataset by combining predictions from multiple trees, reducing overfitting while maintaining high accuracy.

2.4.4 Evaluation

For evaluation models, the following evaluation metrics are used to measure the effectiveness of the models:

2.4.1 Accuracy

Accuracy measures the proportion of correctly classified instances out of the total samples as Equation (5).

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}} \tag{5}$$

While accuracy provides an overall measure of model performance, it may not be sufficient in datasets with previously imbalanced classes. For instance, without handling imbalance, the majority classes ("Neutral Sentiment" or "Negative Sentiment") would dominate the predictions, leading to inflated accuracy scores. After balancing the dataset, accuracy becomes a more reliable indicator as all classes are equally represented during training.

2.4.2 Precision

Precision evaluates the proportion of correctly predicted instances for a specific class out of all instances predicted as Equation (6).

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{6}$$

For this study, precision is crucial for minority classes such as "Threatening Others" or "SARA-Based Content." A high precision score ensures that the model does not falsely classify benign comments as harmful, which is important for maintaining fairness in social media moderation systems.

2.4.3 Recall (Sensitivity)

Recall measures the ability of the model to correctly identify all instances of a specific class as Equation (7).

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{7}$$

Recall is especially important for minority classes in the dataset. For instance, correctly identifying "Threatening Others" comments is critical for detecting potential cybercrimes. A low recall for these classes would mean the model fails to flag harmful content, reducing its effectiveness in real-world applications like enforcing digital laws.

2.4.4 F1-Score

The F1-Score as Equation (8) provides a balanced measure of precision and recall, particularly useful in imbalanced datasets:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

For categories like "Insulting Government" or "SARA-Based Content," where precision and recall may vary, the F1-score offers a single metric to evaluate the trade-off between these two measures. It ensures that the model is not biased toward achieving high precision or recall at the expense of the other.

2.4.5 Macro-Averaged and Weighted-Averaged Metrics

Macro-Averaged Metrics: Compute the average precision, recall, and F1-score across all classes, treating each class equally. This is particularly useful in evaluating the model's performance on minority classes in the balanced dataset. Weighted-Averaged Metrics: Compute averages weighted by the number of samples in each class. This provides a more comprehensive view of the model's overall performance across the dataset.

2.5 Comparative Analysis

The comparative analysis of Naive Bayes, Support Vector Machines (SVM), and Random Forest is conducted to evaluate their performance in classifying social media comments into six sentiment categories. This process helps identify the most suitable model for detecting cybercrimes and harmful content under Indonesia's Information and Electronic Transactions Law (UU ITE).

3. Results and Discussion

3.1 Results

The initial dataset used in this study comprised 7 categories of labels: Neutral Sentiment (0), Positive Sentiment (1), Negative Sentiment (2), Insulting Government (3), Insulting or Defaming Others (4), Threatening Others (5), and SARA-Based Content (6). The preprocessing process for social media comments involves multiple stages to transform raw text into a structured format suitable for machine learning analysis. The initial comment (original) undergoes noise removal, where irrelevant elements such as URLs and special characters are eliminated (no_url and no_punct). Following this, the text is standardized by converting all characters to lowercase (lowercased), ensuring consistency across the dataset. The text is then tokenized, splitting it into individual words or tokens, such as ['gua', 'dah', 'kabor', ...], for easier manipulation. Stopwords, which are common words that do not contribute significantly to the meaning of the text, are removed in the next step (stopwords_removed), resulting in a cleaner representation of meaningful words. Finally, the cleaned tokens are reconstructed into a coherent text format (reconstructed), ready for numerical representation. Figure 3 represents result of pre-processing steps from noise remove to reconstructed step.

```
Preprocessing Results (Example):
original: Gua dah kabor nak bagi gua specific task jangan masuk campur, bajingan'
no_url: Gua dah kabor nak bagi gua specific task jangan masuk campur, bajingan'
no_punct: Gua dah kabor nak bagi gua specific task jangan masuk campur bajingan
lowercased: gua dah kabor nak bagi gua specific task jangan masuk campur bajingan
tokenized: ['gua', 'dah', 'kabor', 'nak', 'bagi', 'gua', 'specific', 'task', 'jangan', 'masuk', 'campur', 'bajingan']
stopwords_removed: ['gua', 'dah', 'kabor', 'nak', 'gua', 'specific', 'task', 'jangan', 'masuk', 'campur', 'bajingan']
reconstructed: gua dah kabor nak gua specific task jangan masuk campur bajingan
```

Fig 3. Result of Pre-Processing

In the final step of pre-processing, the cleaned text is converted into numerical vectors using TF-IDF (Term Frequency-Inverse Document Frequency), which assigns weights to words based on their frequency in a document relative to the entire dataset. Each entry is represented as a sparse matrix, where rows correspond to comments and columns correspond to unique words, weighted by their importance. For instance, the TF-IDF matrix shows the contribution of specific words to the overall representation of a comment, enabling algorithms to process textual data numerically. Result of Numerical Representation is represented in Fig 4.

```
Numerical Representation (TF-IDF Example):
  (0, 3912)     0.22686900566409818
  (0, 6)        0.20894032604527016
  (0, 2328)     0.17530667071840123
  (0, 1422)     0.20088884507319074
  (0, 220)      0.23608561430928465
  (0, 2163)     0.3038411096995637
  (0, 1821)     0.17374355680917622
  (0, 2026)     0.23608561430928465
  (0, 4568)     0.22686900566409818
  (0, 4800)     0.20088884507319074
  (0, 4790)     0.4461910680276618
  (0, 2064)     0.23608561430928465
  (0, 123)      0.22686900566409818
  (0, 2414)     0.21972004859968164
  (0, 1783)     0.1991566675525899
  (0, 741)      0.24907569460473838
  (0, 4833)     0.14064835919938104
  (0, 1296)     0.17868215613255053
```

Fig 4. Result of Numerical Representation

The dataset was highly imbalanced, with the majority of the samples belonging to Neutral Sentiment (0) and Positive Sentiment (1), while minority classes such as Threatening Others (5) and SARA-Based Content (6) were underrepresented. This imbalance posed challenges for the machine learning models, as they tend to favor majority classes, leading to poor performance in detecting minority class violations. To address the research objective of identifying potential cybercrimes under Indonesia's Information and Electronic Transactions Law (UU ITE), only 4 relevant categories were retained during preprocessing based on Table 2.

Table 2. Relevant Categories for indentifying potential cybercrime under UU ITE

Insulting Government (3)	Comments insulting the government or public institutions.
Insulting or Defaming Others (4)	Comments defaming or insulting individuals.
Threatening Others (5)	Comments containing direct threats to individuals or groups.
SARA-Based Content (6)	Comments targeting ethnicity, religion, race, or intergroup relations.

Categories such as Neutral Sentiment (0) and Positive Sentiment (1) were excluded because they do not represent behaviors directly associated with legal violations. As a result, the dataset size was reduced significantly, from 12,647 samples to 4,530, containing only comments related to these 5 categories. This filtering ensured that the analysis remained focused on identifying harmful or illegal behaviors. Even after filtering, the dataset remained imbalanced, with categories like Threatening Others (5) and SARA-Based Content (6) still containing far fewer samples compared to Negative Sentiment (2). For instance, the class distribution before SMOTE was as Table 3.

Table 3. Class distribution before SMOTE

Negative Sentiment (2)	3355 samples
Insulting Government (3)	120 samples
Insulting or Defaming Others (4)	78 samples
Threatening Others (5)	44 samples
SARA-Based Content (6)	24 samples

This imbalance would negatively affect the model's ability to detect minority class instances effectively, as the models might overfit to majority classes (Insulting Government) and ignore the minority ones. To resolve this, Synthetic Minority Oversampling Technique (SMOTE) was applied to the training data. SMOTE works by generating synthetic samples for the minority classes, balancing the dataset. After applying SMOTE, each class contained an equal number of samples. Class distribution after SMOTE is shown in Table 4.

Table 4. Class distribution after SMOTE

Negative Sentiment (2)	3355 samples
Insulting Government (3)	3355 samples
Insulting or Defaming Others (4)	3355 samples
Threatening Others (5)	3355 samples
SARA-Based Content (6)	3355 samples

After balancing, the dataset was split into training (80%) and testing (20%) sets, resulting in 3,624 samples for training and 906 samples for testing. The balancing process, performed using SMOTE, was applied only to the training set. This ensured an equal number of samples for each class in the training set, with 3,355 samples per class, effectively addressing the issue of class imbalance. The testing set, however, retained its original class distribution to provide an accurate evaluation of the model's performance on real-world, imbalanced data.

This research compared the performance of Naive Bayes, Support Vector Machines (SVM), and Random Forest algorithms in identifying and classifying cybercrimes in social media comments, focusing on five sentiment categories: Negative Sentiment, Insulting Government, Insulting or Defaming Others, Threatening Others, and SARA-Based Content. The results highlight that all algorithms excelled at identifying the dominant class (Negative Sentiment) while facing substantial challenges with minority classes, underscoring the impact of dataset imbalance.

Table 5. Classification Report of Naive Bayes

Class	Precision	Recall	F1-Score	Support
Negative Sentiment (2)	0.96	0.75	0.84	833
Insulting Government (3)	0.19	0.43	0.27	14
Insulting or Defaming Others (4)	0.11	0.44	0.18	32
Threatening Others (5)	0.09	0.40	0.15	5
SARA-Based Content (6)	0.13	0.45	0.21	22
Accuracy			0.73	906
Macro Average	0.30	0.49	0.33	906
Weighted Average	0.89	0.73	0.79	906

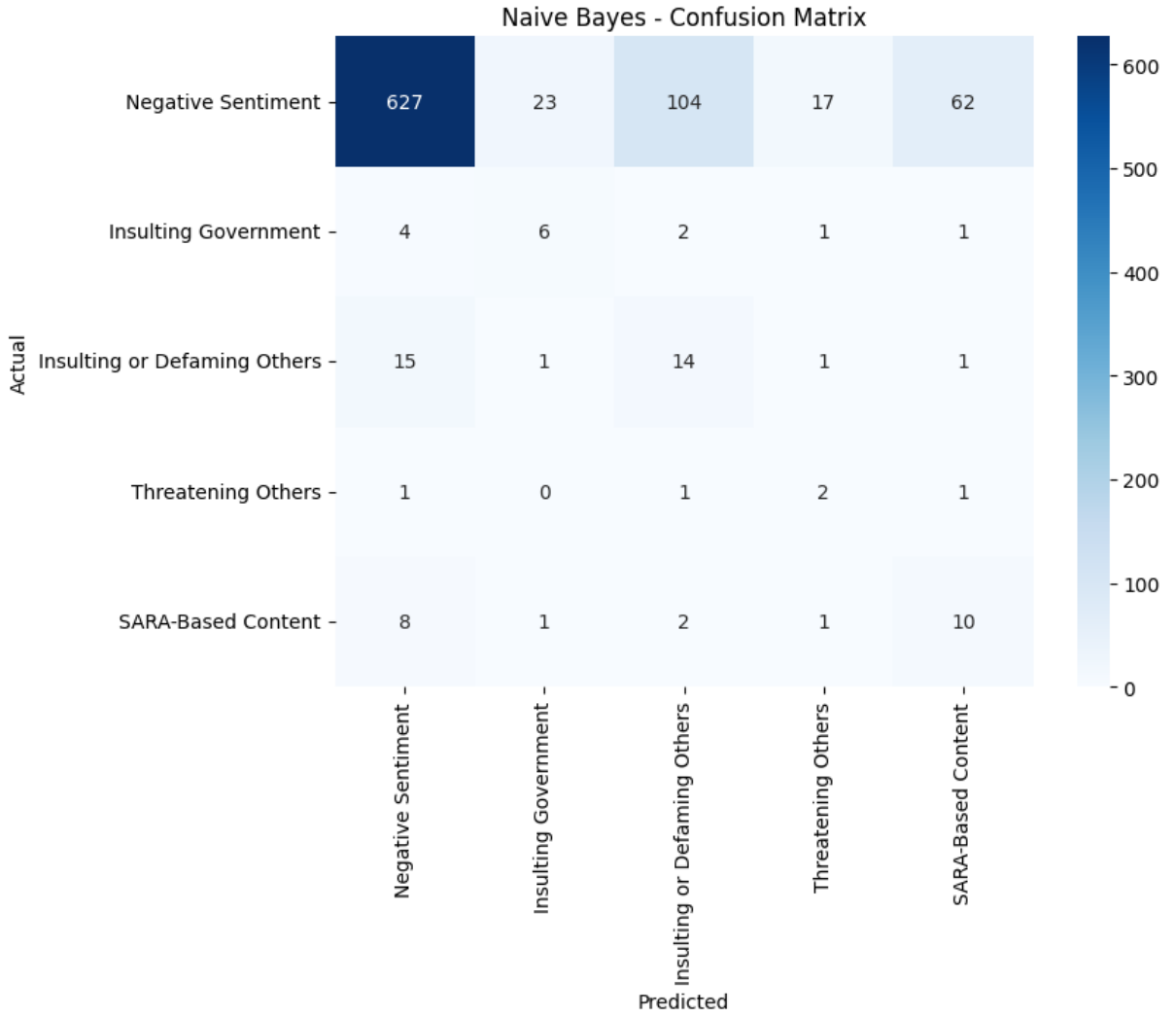

Fig 5. Confusion Matrix of Naive Bayes

Based on Table 5, Naive Bayes achieved an accuracy of 73%, showing strong performance for the Negative Sentiment class, with a precision of 0.96, recall of 0.75, and an F1-score of 0.84. However, for minority classes, its performance dropped significantly. For example, the F1-score for Insulting Government was only 0.27, and for SARA-Based Content, it was 0.21. The macro-average precision and F1-scores were 0.30 and 0.33, respectively, reflecting the algorithm's limited ability to generalize across all classes. Confusion Matrix of Naive Bayes is represented in Figure 5.

Table 6. Classification Report of SVM

Class	Precision	Recall	F1-Score	Support
Negative Sentiment (2)	0.93	0.94	0.93	833
Insulting Government (3)	0.56	0.36	0.43	14
Insulting or Defaming Others (4)	0.11	0.16	0.13	32
Threatening Others (5)	0.00	0.00	0.00	5
SARA-Based Content (6)	0.25	0.09	0.13	22
Accuracy			0.88	906
Macro Average	0.37	0.31	0.33	906
Weighted Average	0.87	0.88	0.87	906

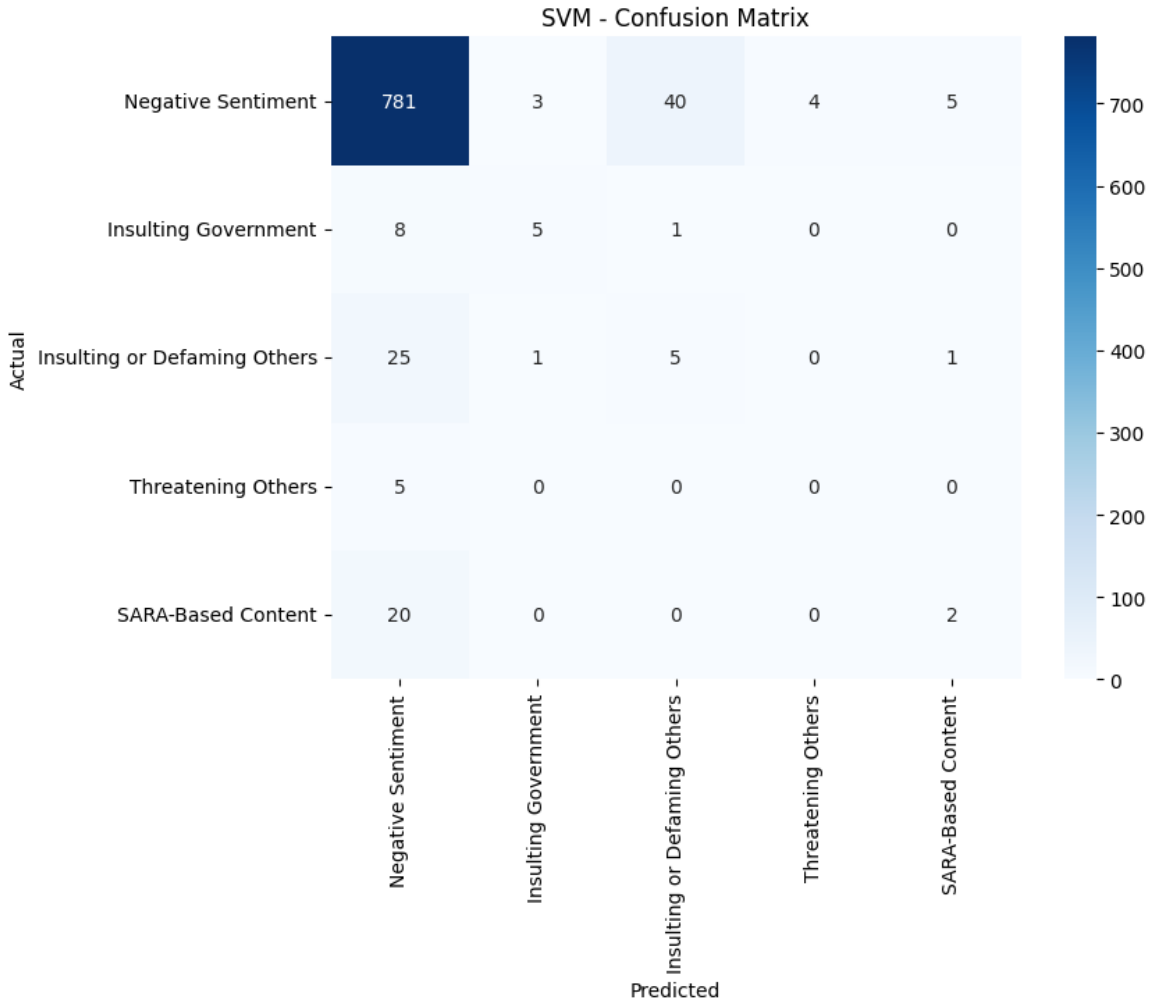

Fig 6. Confusion Matrix of SVM

SVM outperformed Naive Bayes with an overall accuracy of 88% based on Table 6. It achieved excellent results for Negative Sentiment, with an F1-score of 0.93, but struggled with minority classes. For Insulting Government, it achieved a precision of 0.56 and an F1-score of 0.43. However, for Threatening Others and SARA-Based Content, SVM recorded F1-scores of 0.00 and 0.13, respectively, showing its limitations in detecting underrepresented classes. The macro-average F1-score of 0.33 indicated moderate improvement compared to Naive Bayes. Confusion Matrix of SVM is shown in Figure 6.

Table 7. Classification Report of Random Forest

Class	Precision	Recall	F1-Score	Support
Negative Sentiment (2)	0.93	0.98	0.95	833
Insulting Government (3)	0.50	0.29	0.36	14
Insulting or Defaming Others (4)	0.38	0.09	0.15	32
Threatening Others (5)	0.09	0.00	0.00	5
SARA-Based Content (6)	0.00	0.00	0.00	22
Accuracy			0.91	906
Macro Average	0.36	0.27	0.29	906
Weighted Average	0.87	0.91	0.89	906

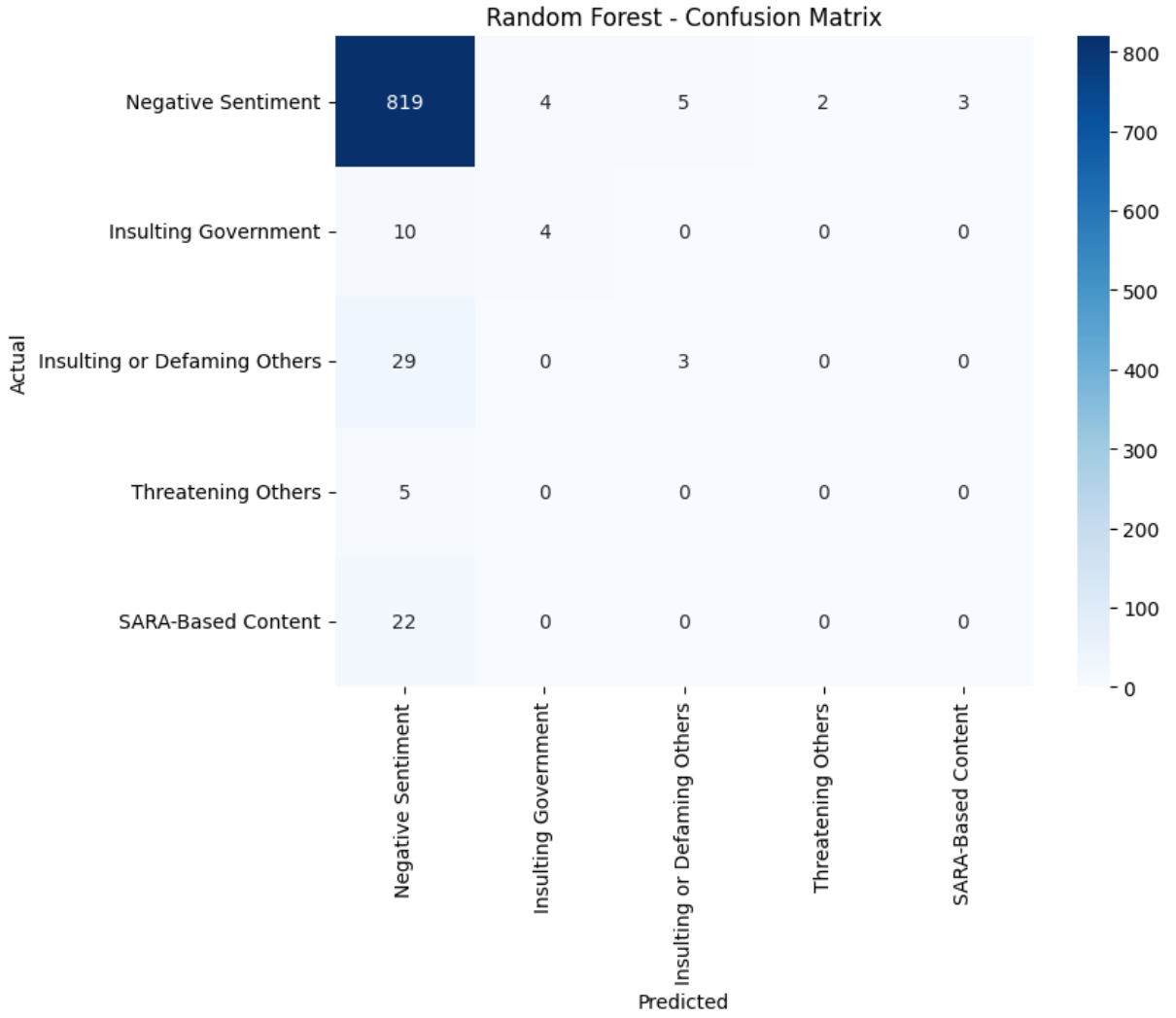

Fig 7. Confusion Matrix of Random Forest

Based on Table 7 dan Figure 7, Random Forest emerged as the top-performing algorithm, with an overall accuracy of 91%. It achieved exceptional results for Negative Sentiment, with a precision of 0.93, recall of 0.98, and an F1-score of 0.95. For minority classes, Random Forest demonstrated better results than Naive Bayes and SVM, achieving an F1-score of 0.36 for Insulting Government and 0.15 for SARA-Based Content. However, it failed to detect Threatening Others, resulting in an F1-score of 0.00 for that class. The macro-average F1-score of 0.29 reflects the persistent difficulty in achieving balance across all classes.

3.2 Discussion

The findings of this study reveal the strengths and limitations of Naive Bayes, Support Vector Machines (SVM), and Random Forest in detecting and classifying harmful social media comments related to cybercrimes. Based on Figure 8, a persistent challenge faced by all algorithms was the handling of class imbalance, as they performed well for the dominant class (Negative Sentiment) but struggled with the minority classes (Insulting Government, Insulting or Defaming Others, Threatening Others, and SARA-Based Content).

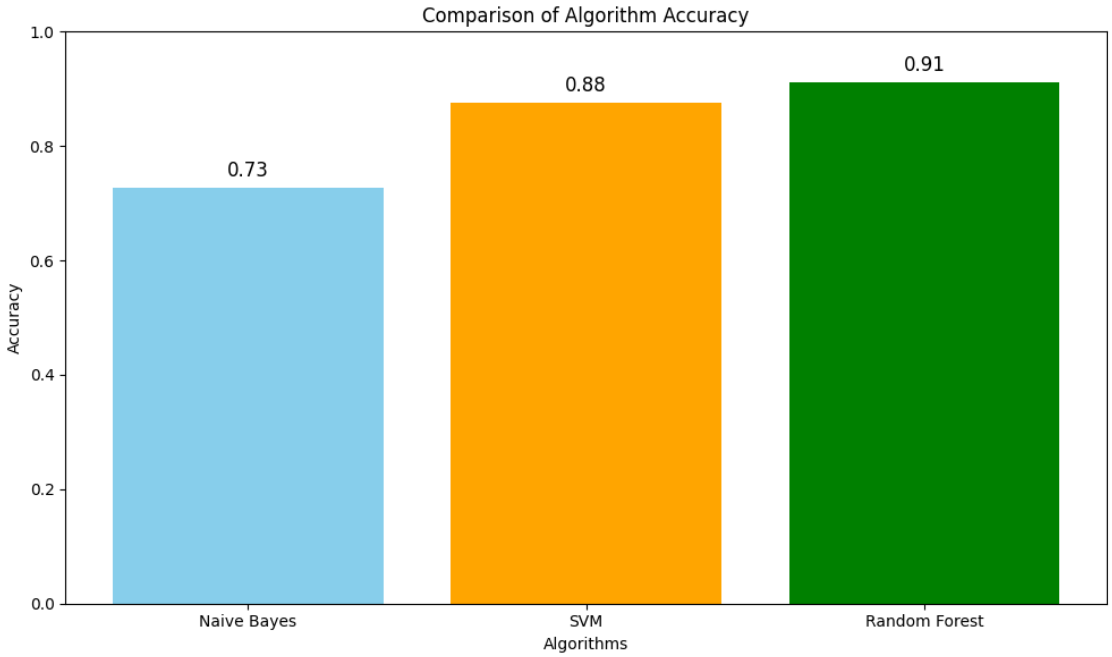

Fig 7. Comparison of Algoritm Accuracy

Naive Bayes, known for its computational efficiency, achieved an accuracy of 73%. While it excelled in identifying Negative Sentiment (F1-score: 0.84), its performance dropped significantly for minority classes, such as Insulting Government (F1-score: 0.27) and SARA-Based Content (F1-score: 0.21). These results highlight its reliance on the assumption of feature independence, which is inadequate for capturing the complexity of imbalanced textual data. Although it remains a viable option for scenarios requiring rapid predictions, Naive Bayes struggles to generalize across minority classes, as indicated by its macro-average F1-score of 0.33. SVM demonstrated a more balanced performance, achieving 88% accuracy and a macro-average F1-score of 0.33. It delivered strong results for Negative Sentiment (F1-score: 0.93) and improved performance for Insulting Government (F1-score: 0.43) compared to Naive Bayes. However, SVM faced significant challenges with underrepresented classes such as Threatening Others, where it failed to detect any instances. Additionally, its moderate F1-score of 0.13 for SARA-Based Content highlights its limitations. Despite its ability to handle complex textual features, SVM's computational demands and sensitivity to parameter tuning make it less practical for real-time or resource-constrained applications. Random Forest emerged as the top-performing algorithm, achieving the highest accuracy (91%) and consistently outperforming Naive Bayes and SVM on most metrics. It excelled in classifying Negative Sentiment (F1-score: 0.95) and achieved the best results for Insulting Government (F1-score: 0.36). However, like the other models, it failed to effectively classify Threatening Others and SARA-Based Content, with F1-scores of 0.00 and 0.00, respectively. The macro-average F1-score of 0.29 underscores the persistent difficulty of achieving balanced performance across all classes, even with ensemble learning techniques.

In summary, Random Forest provided the most consistent performance across classes, particularly for the dominant category, while SVM offered a viable alternative with moderate improvements for certain minority classes. Naive Bayes, though computationally efficient, demonstrated significant limitations with imbalanced datasets. These results underscore the challenges of detecting cybercrimes in diverse and imbalanced social media datasets. In this study, hyperparameters were not selected based on scientific or domain-specific reasoning because the primary objective was to evaluate the baseline performance of the models in classifying harmful content in social media comments. This approach offers an initial understanding of the strengths and weaknesses of the models without the added complexity of hyperparameter optimization. As noted in [25], using default or basic optimization techniques is acceptable in early research stages to establish foundational insights.

In future research, hyperparameter selection will be conducted using a more domain-informed and theoretical approach to enhance model performance and adaptability to the data. As suggested in [26], strategies such as data-driven hyperparameter tuning, analyzing the impact of parameters on minority classes, and leveraging advanced optimization techniques like Bayesian Optimization or Genetic Algorithms will be explored. Additionally, future studies will consider integrating contextual feature engineering and transfer learning to further improve the model's generalizability and ability to handle imbalanced data effectively.

4. Conclusion

This study evaluated the performance of Naive Bayes, Support Vector Machines (SVM), and Random Forest in identifying and classifying harmful content in social media comments, focusing on defamation, threats, and SARA-based content. The results showed that Random Forest achieved the highest accuracy (91%) and excelled in classifying dominant classes like Negative Sentiment (F1-score: 0.95) and moderately represented classes such as Insulting Government (F1-score: 0.36). However, it struggled with underrepresented classes like Threatening Others and SARA-Based Content, where it achieved F1-scores of 0.00. SVM delivered strong results with an accuracy of 88%, excelling in handling dominant classes (F1-score: 0.93) and performing moderately well for Insulting Government (F1-score: 0.43). However, like Random Forest, it faced challenges with minority classes and was limited by its computational complexity. Naive Bayes had the lowest accuracy (73%), performing well for Negative Sentiment (F1-score: 0.84) but showing significant limitations with minority classes due to its reliance on simplistic assumptions. Overall, Random Forest emerged as the most effective algorithm for this task, but all algorithms struggled with minority class detection. Addressing class imbalance remains critical, and future work should explore advanced techniques like data augmentation, contextual feature engineering, and hybrid models to improve classification performance and ensure balanced detection of harmful content, supporting legal frameworks such as Indonesia's UU ITE.

5. Acknowledgment

The authors would like to express their sincere gratitude to the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia for their generous funding and support in the realization of this research through the PDP Scheme. This financial support has been instrumental in the successful completion of this study on the application of machine learning for identifying cybercrimes in social media comments, particularly in fostering a deeper understanding of digital citizenship and supporting the enforcement of Indonesia's UU ITE. The authors are deeply appreciative of the opportunity to contribute to the development of innovative solutions that promote ethical and responsible online behavior in the digital era.

7. References

[1] A. Sharma and U. Ghose, “Sentimental Analysis of Twitter Data with respect to General Elections in India,” in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 325–334. doi: 10.1016/j.procs.2020.06.038.

[2] S. Zervoudakis, E. Marakakis, H. Kondylakis, and S. Goumas, “OpinionMine: A Bayesian-based framework for opinion mining using Twitter Data,” *Mach. Learn. with Appl.*, vol. 3, p. 100018, Mar. 2021, doi: 10.1016/j.mlwa.2020.100018.

[3] M. T. Hasan, M. A. E. Hossain, M. S. H. Mukta, A. Akter, M. Ahmed, and S. Islam, “A Review on Deep-Learning-Based Cyberbullying Detection,” *Futur. Internet*, vol. 15, no. 5, pp. 1–47, 2023, doi: 10.3390/fi15050179.

[4] S. J. Lesmana, I. Sofia, and F. Felina, “Law Enforcement in Efforts to Combat Cyber Crime in Indonesia Building Future Digital Security,” *Int. J. Law Rev. …*, vol. 1, no. 3, pp. 120–128, 2023, [Online]. Available: https://www.ijems.id/index.php/ijlrsa/article/view/90%0Ahttps://www.ijems.id/index.php/ijlrsa/article/download/90/77

[5] Waluyadi, “Law Enforcement Against Cyber Crimes in Indonesia : Analysis of The Role of The

Ite Law in Handling Cyber Crimes," *Indones. Cyber Law Rev.*, vol. 1, no. 1, pp. 38–46, 2024.

[6] D. Sukardi, "Legal Protection Against Cyber Crime Threats for Business Economics 4.0," *Leg. Br.*, vol. 11, no. 4, pp. 2227–2235, 2022, doi: 10.35335/legal.

[7] B. G. Bokolo and Q. Liu, "Artificial Intelligence in Social Media Forensics: A Comprehensive Survey and Analysis," *Electron.*, vol. 13, no. 9, 2024, doi: 10.3390/electronics13091671.

[8] F. R. S. Nascimento, G. D. C. Cavalcanti, and M. Da Costa-Abreu, "Exploring Automatic Hate Speech Detection on Social Media: A Focus on Content-Based Analysis," *SAGE Open*, vol. 13, no. 2, pp. 1–19, 2023, doi: 10.1177/21582440231181311.

[9] G. Kovács, P. Alonso, and R. Saini, "Challenges of Hate Speech Detection in Social Media: Data Scarcity, and Leveraging External Resources," *SN Comput. Sci.*, vol. 2, no. 2, pp. 1–15, 2021, doi: 10.1007/s42979-021-00457-3.

[10] A. Muneer and S. M. Fati, "A comparative analysis of machine learning techniques for cyberbullying detection on twitter," *Futur. Internet*, vol. 12, no. 11, pp. 1–21, 2020, doi: 10.3390/fi12110187.

[11] I. S. Arfan, S. Fauziah, and I. Nawangsih, "Sentiment Analyst of Cyber Bullying in X Using Naïve Bayes Algorithm Analisa Sentimen Terhadap Cyber Bullying di X Menggunakan Algoritma Naïve Bayes," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. October, pp. 1411–1419, 2024.

[12] A. Muzakir, H. Syaputra, and F. Panjaitan, "A Comparative Analysis of Classification Algorithms for Cyberbullying Crime Detection: An Experimental Study of Twitter Social Media in Indonesia," *Sci. J. Informatics*, vol. 9, no. 2, pp. 133–138, 2022, doi: 10.15294/sji.v9i2.35149.

[13] N. Novalita, A. Herdiani, I. Lukmana, and D. Puspandari, "Cyberbullying identification on twitter using random forest classifier," *J. Phys. Conf. Ser.*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012029.

[14] D. Hidayat, H. Firmanda, and M. H. Wafi, "Analysis of Hate Speech in the Perspective of Changes to the Electronic Information and Transaction Law," *Fiat Justisia J. Ilmu Huk.*, vol. 18, no. 1, pp. 31–48, 2024, doi: 10.25041/fiatjustisia.v18no1.3146.

[15] F. Isnawan, "TINJAUAN HUKUM PIDANA TENTANG FENOMENA CYBERBULLYING YANG DILAKUKAN OLEH REMAJA," *J. Interpret. Huk.*, vol. 4, no. 1, pp. 145–163, 2023.

[16] J. Guo *et al.*, "A Deep Look into neural ranking models for information retrieval," *Inf. Process. Manag.*, vol. 57, no. 6, p. 102067, 2020, doi: 10.1016/j.ipm.2019.102067.

[17] M. A. Rofiqi, A. C. Fauzan, A. P. Agustin, and A. A. Saputra, "Implementasi Term-Frequency Inverse Document Frequency (TF- IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query," *J. Comput. Sci. Appl. Informatics*, vol. 1, no. 2, pp. 58–64, 2019.

[18] J. Pardede and D. P. Pamungkas, "The Impact of Balanced Data Techniques on Classification Model Performance," *Sci. J. Informatics*, vol. 11, no. 2, pp. 401–412, 2023, doi: 10.15294/sji.v11i2.3649.

[19] A. M. Halim, M. Dwifebri, and F. Nhita, "Handling Imbalanced Data Sets Using SMOTE and ADASYN to Improve Classification Performance of Ecoli Data Sets," *Build. Informatics, Technol. Sci.*, vol. 5, no. 1, pp. 246–253, 2023, doi: 10.47065/bits.v5i1.3647.

[20] K. Hikmah and A. C. Fauzan, "Sentiment Analysis of Vaccine Booster during Covid-19: Indonesian Netizen Perspective Based on Twitter Dataset," *J. Teknol. Komput. dan Sist. Inf.*, vol. 5, no. 2, pp. 102–106, 2022.

[21] A. C. Fauzan and K. Hikmah, "Implementasi Algoritma Naive Bayes Dalam Analisis Polarisasi Opini Masyarakat Terkait Vaksin Covid-19," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 7, no. 2, pp. 122–128, 2022, doi: 10.36341/rabit.v7i2.2403.

[22] D. Rizki Surya Pratama, T. A. Munandar, and K. Fadhilla Ramdhania, "Multinomial Naive Bayes Algorithm for Indonesian language Sentiment Classification Related to Jakarta International Stadium (JIS)," *Int. J. Inf. Technol. Comput. Sci. Appl.*, vol. 2, no. 1, pp. 12–22, 2024, doi: 10.58776/ijitcsa.v2i1.118.

[23] R. Mukarramah, D. Atmajaya, and L. B. Ilmawan, "Performance comparison of support vector machine (SVM) with linear kernel and polynomial kernel for multiclass sentiment analysis on twitter," *Ilk. J. Ilm.*, vol. 13, no. 2, pp. 168–174, 2021, doi: 10.33096/ilkom.v13i2.851.168-174.

[24] P. Handayani and A. Charis Fauzan, "Machine Learning Klasifikasi Status Gizi Balita Menggunakan Algoritma Random Forest," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 6, pp. 3064–3072, 2024, doi: 10.30865/klik.v4i6.1909.

[25] Y. Bergstra, James Bengio, "Random Search for Hyper-Parameter Optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012.

[26] M. Feurer and F. Hutter, *Hyperparameter Optimization*. Springer, 2019.