Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 9, No. 1, March 2026, pp. 140 – 152
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Comparative Study of Machine Learning Methods for Sentiment Analysis of TikTok Comments Related to Cyberbullying

[1*]Celestina Florecita Mariwy, [2]Lorna Yertas Baisa, [3]Andreas Leonardo Sumendap
[1,2,3]Department of Informatics Engineering, Faculty of Engineering, University of Papua, Indonesia
Email: [1]fmariwy@gmail.com, [2]L.baisa@unipa.ac.id, [3]al.sumendap@unipa.ac.id

Article Info

Article history:
Received Jan 19th, 2026
Revised Feb 31th, 2026
Accepted Mar 11th, 2026

Keyword:
Cyber Bullying
K-Nearest Neighbors
Naive Bayes
Sentiment Analysis
Support Vector Machine

ABSTRACT

The rapid growth of internet use in Indonesia has contributed to the rise of cyberbullying on TikTok, increasing the importance of automated sentiment analysis for digital safety. This study compares the performance of Support Vector Machine, K-Nearest Neighbors, and Naive Bayes for sentiment classification in TikTok comments related to cyberbullying. The dataset was collected via web scraping and processed through several preprocessing stages, yielding 7,900 unique comments. Sentiment labeling used a lexicon-based approach, and the data were split into training and testing sets with an 80:20 ratio. Results show that 34.18% of comments were negative, indicating a notable level of harmful content. Among the three models, Support Vector Machine performed best with an accuracy of 91.5%, followed by Naive Bayes at 82.8% and K-Nearest Neighbors at 80.8%. These findings suggest Support Vector Machine is the most effective method for sentiment classification in this context and offer a useful reference for developing more accurate content moderation systems on social media.

Corresponding Author:
Celestina Florecita Mariwy,
Department of Informatics Engineering,
Faculty of Engineering, University of Papua,
Gunung Salju Amban St., Manokwari, West Papua, Indonesia, Postal Code 98314.
Email: fmariwy@gmail.com

DOI: http://dx.doi.org/10.24014/ijaidm.v9i1.39183

1. INTRODUCTION

The increasing internet penetration in Indonesia, projected to reach 79.5% by 2024, brings both opportunities and risks, particularly for users' psychological well-being. As social media becomes deeply embedded in daily life, the need for stronger digital literacy and safer online interaction has become increasingly urgent to reduce toxic behavior and cyberbullying in digital spaces [1]. Although Indonesia has more than 143 million social media users, the country's Digital Society Index (IDSI) 2025 score of 44.53 still indicates a limited level of public digital literacy, which contributes to the persistence of harmful online communication [2].

Among social media platforms, TikTok has become one of the most influential in Indonesia, with approximately 108 million users. Its algorithm-driven content distribution, interactive features, and anonymous comment environment make it highly vulnerable to the spread of offensive language and cyberbullying. As a result, identifying harmful sentiment in TikTok comments is increasingly important for supporting safer online engagement and content moderation [3]. Similar concerns have been reported in other digital environments, including online gaming communities, indicating that cyberbullying is a broader, persistent issue across interactive platforms [3].

Because TikTok generates an enormous number of comments daily, manual moderation can no longer detect harmful content at scale. Automated sentiment analysis therefore provides a practical way to classify comments as positive, neutral, or negative [4]. However, this task remains challenging in the Indonesian context because TikTok comments often contain slang, abbreviations, informal spelling, and local expressions

that impair classification performance [5]. Recent studies have examined sentiment analysis and cyberbullying detection on social media using machine learning methods [6],[7], but most still focus on general platforms, different languages, or limited model comparisons. In particular, comparative studies that specifically evaluate classical machine learning algorithms on Indonesian TikTok comments related to cyberbullying remain limited.

This study addresses that gap by comparing Support Vector Machine, K-Nearest Neighbors, and Naive Bayes for sentiment analysis of Indonesian TikTok comments related to cyberbullying. The novelty of this research lies in its focus on TikTok-specific comments, its emphasis on cyberbullying-related discourse in Indonesian, and its direct comparison of three widely used machine learning methods under the same preprocessing and evaluation settings. The main objective is to identify the most effective model for sentiment classification in this context. The findings are expected to contribute to the development of more accurate automated moderation systems and provide a practical reference for future research on cyberbullying detection in Indonesian social media. The remainder of this article is structured as follows: Section 2 describes the research methodology, Section 3 presents the results and analysis, and Section 4 concludes the study and suggests directions for future work.

2. RESEARCH METHOD

This study employs a comparative experimental design to evaluate the performance of three classical machine learning algorithms: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naive Bayes. These algorithms were chosen because they are widely used in text classification and sentiment analysis studies [8], [9]. The research workflow consists of four main stages: data collection; preprocessing and labeling; feature extraction and model implementation; and performance evaluation. The overall research framework is illustrated in Figure 1.

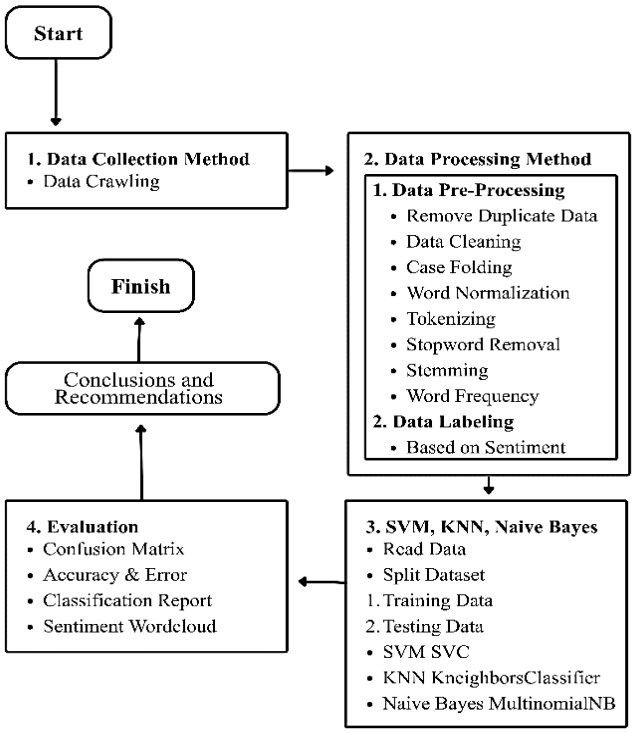

Figure 1. Research Phase

2.1. Data Collection Method

TikTok, a platform characterized by rapid interaction and massive content algorithms, offers a broad space for users to discuss social phenomena. However, this environment also facilitates the emergence of aggressive behaviors and cyberbullying-related discussions. To effectively capture these dynamics for sentiment analysis, data was meticulously extracted from TikTok comments.

The selection of TikTok videos for data scraping was strategically focused on content that had garnered significant public attention and was identified as potentially triggering cyberbullying discussions, including viral news or controversial topics known to elicit strong reactions in comment sections. This approach aimed to ensure the collected comments were representative of actual cyberbullying-related discourse

on the platform. Data extraction was performed on December 20, 2025, using web crawling techniques via the TikTok Comment Scraper tool from Apify (https://apify.com/quacker/tiktok-comments-scraper), technical procedure resulted in the collection of 10,188 comment entries. Subsequently, these raw comments were subjected to an initial filtering process to remove irrelevant entries (e.g., advertisements, duplicate comments, or comments entirely unrelated to the video's content or the broader discussion of cyberbullying). The filtered data were then converted into Excel format to facilitate subsequent data cleaning and algorithm testing. This approach is consistent with efforts to build specialized datasets for social media analysis [10]

2.2. Data Preprocessing Method

Because comments from social media often contain noise, informal expressions, slang, and inconsistent spelling, preprocessing is an essential step before model training [11]. In this study, the collected text data underwent the following preprocessing stages:

1. Cleaning: URLs, hashtags, mentions, emojis, punctuation, special characters, and repeated spaces were removed to reduce noise.
2. Case Folding: All text was converted to lowercase to ensure consistency in word representation.
3. Normalization: Informal words, abbreviations, slang, and local expressions were normalized into standard Indonesian forms using a manually constructed dictionary based on frequently occurring TikTok expressions.
4. Tokenization: Each comment was split into individual words to facilitate feature extraction.
5. Stopword Removal: Common function words with low semantic value were removed using an Indonesian stopword list from the NLTK library.
6. Stemming/Lemmatization: Words were reduced to their base forms to improve vocabulary consistency across similar terms.

After preprocessing, sentiment labeling was performed using a lexicon-based approach. A custom Indonesian sentiment lexicon was used to assign polarity labels to each comment: positive, neutral, or negative. To improve label reliability, a subset of the labeled data was manually reviewed by two independent annotators. Agreement between annotators was measured with Cohen's Kappa, and disagreements were resolved through discussion. This verification step aimed to reduce errors caused by sarcasm, ambiguous expressions, and slang-heavy language, which are common in TikTok comments.

2.3. Classification of SVM, KNN, and Naive Bayes

This study compares three supervised machine learning algorithms Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naive Bayes to determine the most effective model for sentiment classification of TikTok comments. These algorithms were chosen because they represent different classification paradigms and have been widely applied in text mining and sentiment analysis research. SVM is a margin-based classifier effective at handling high-dimensional feature spaces [12], [13]. KNN is an instance based method that classifies data according to the nearest training samples[14]. Naive Bayes is a probabilistic classifier that estimates class membership using Bayes' theorem under conditional independence assumptions[15]. For SVM, the decision function is defined as Equation 1.

$$f(x) = \text{sign}\,(w^T x + b) \tag{1}$$

where x is the input vector, w is the weight vector, and b is the bias term. The optimization objective of SVM is Equation 2.

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \xi_i \tag{2}$$

with constraints, Equation 3.

$$y_i\,(w^T x_i + b) \geq 1 - \xi_i,\ \ \xi_i \geq 0 \tag{3}$$

This formulation allows SVM to maximize the margin while controlling classification errors through the penalty parameter C. For KNN, the similarity between a test sample and training samples is commonly measured using Euclidean distance as Equation 4.

$$d(x, x_i) = \sqrt{\sum_{j=1}^{m} \left(x_j - x_{ij}\right)^2} \tag{4}$$

The predicted class is then determined by majority voting among the KNN as in Equation 5.

$$\hat{y} = \arg\max_{c \in C} \sum_{i \in N_k(x)} I(y_i = c) \quad (5)$$

Where $N_k(x)$ denotes the set of k nearest neighbors of x, C is the set of possible classes, and $\mathbb{I}$ is the indicator function. KNN is suitable for sentiment classification when local similarity patterns are informative. For Naive Bayes, the posterior probability of class c given a document x is expressed as Equation 6.

$$P(c|x) = \frac{P(c)\Pi_{j=1}^{m} P(x_j|c)}{P(x)} \quad (6)$$

Since $P(x)$ is constant across all classes, the predicted class is obtained as Equation 7.

$$\hat{c} = \arg\max_{c \in C} P(c)\Pi_{j=1}^{m} P(x_j|c) \quad (7)$$

This classifier is computationally efficient and widely applied in text mining because it handles sparse feature spaces effectively.

Overall, the comparison of SVM, KNN, and Naive Bayes provides a balanced evaluation of margin-based, distance-based, and probabilistic approaches for classifying Indonesian TikTok comments, which are characterized by slang, abbreviations, and informal language. The results of this comparison are expected to support the selection of an appropriate model for sentiment analysis and cyberbullying-related content moderation.

2.4. Evaluation

The dataset was split into training and testing sets using an 80:20 ratio [16]. To ensure a fair comparison, all models were trained and evaluated on the same data partition and preprocessing pipeline. Model performance was assessed using the following metrics as Accuracy, Precision, Recall, F1-score, and Confusion Matrix.

These metrics measured each model's ability to correctly classify sentiment labels and to reveal misclassification patterns. Accuracy indicates overall classification performance, while precision, recall, and F1-score offer detailed insights into class-level prediction quality, particularly for imbalanced sentiment distributions. The confusion matrix shows how many positive, neutral, and negative comments were correctly or incorrectly classified. This evaluation strategy enables identification of the most effective model for sentiment classification of TikTok comments and assessment of its suitability for cyberbullying-related text analysis.

3. RESULTS AND ANALYSIS

3.1. Data Collection

3.1.1. TikTok Data Scraping

Three videos from the accounts @VISI.NEWS, @sctv_, and @maveth were used as the data sources in this research. The comments were collected from these videos through Apify.com and then converted into CSV format for the next processing stage.

Table 1. Results of Comment Scraping

No	comment_text	number_of_likes	number_of_replies	username
1	PEMBULLY HARUS MENDAPATKAN SANKSI HUKUM AGAR JERA ATAU HARUS MENDAPATKAN SANKSI SOSIAL!!	3153	17	giddy_panda
2	TOLONG BULLYING INI DI TINDAKLANJUTI DENGAN SERIUSSS PLISSSS, ZAMAN SEKARANG MEMBULLY TEMEN YANG LAIN MALAH PADA NGERASA BANGGA!!!	118	1	rgnbalqisss
3	lagii lagii kasus pembullyan😢	1385	4	rtnraeni_
4	Stop!! kalian sedang melakukan cyber bullying.	5	0	fizzzzzrzzz
5	aku dulu juga di bully fisik 😭 Krena item dekil jelek juga 😭 kena bully 1 laki2 habis2an, ingat sampai saat ini 😭 pdhal kejadian udah beberapa tahun lalu	304	0	langit.senjaa04
6	kakk emng ya bekas bully tuh susah lupa nya masih tengngiang "sampe sekarang	71	0	kepoamat2.3
7	kek rada Gedeg nya masih aja gasi	0	0	itsmine.12

No	comment_text	number_of_likes	number_of_replies	username
8	betul kak, kdang ada rasa trauma juga 😭 tiap ketemu orang nya aku kaya sakit hati, dan GK ikhlas banget	29	0	zzhhrtl17
9	kak kita sama kok smngt ya	1	0	eniwidawati2
10	rill ingat seumur hidup😭 benci bgt sm yg suka bully	2	0	quennn354
…..	…..	…..	…..	…..
10188	ada yg ngebully kok dimasa bodohi	4	0	l_nd13

Table 1 shows the results of scraping comments from the TikTok posts of the accounts @VISI.NEWS, @sctv_, and @maveth, totaling 10,188 comments. The video from the @VISI.NEWS account generated 1,877 comments, the @sctv_ account produced 4,728 comments, and the video from the @maveth account generated 3,583 comments. Table 2 presents the scraping results of the comments in tabular form.

Table 2. Comment Results Data

TikTok Account	Video Theme	Data Quantity
@VISI.NEWS	Bullying cases have occurred again in the school environment	1,877 Comment
@sctv_	Stop making hurtful comments! Because every word can be painful	4,728 Comment
@maveth	Cyberbullying can happen to anybody n anywhere, rest in peace Amanda Todd I wish you knew how many people out there love and still care about you till this day and you will never be forgotten	3,583 Comment
	Total Data	10,188 Comment

3.2. Data Pre-processing

Data preprocessing using text mining methods to extract valuable insights is carried out through two main stages. First, raw data is organized in a structured manner, and second, the data is processed to produce useful information. The preprocessing stages include:

3.2.1. Remove Duplicate Data

Removing duplicate data helps maintain the integrity and quality of the analysis. Comments with identical text are eliminated to avoid redundancy and bias, especially in sentiment analysis, which requires unique data. This step aims to prevent distortion of the results, ensuring more valid and objective conclusions.

Table 3. Result Of Removing Duplicate Data

Category	Total Data
Scraping Result	10,188
After Remove Duplicate	7,900

Table 3 shows that a total of 2,288 duplicate data entries were removed from the initial 10,188 comments, leaving 7,900 clean data entries for further analysis to obtain more accurate and reliable insights.

3.2.2. Text Preprocessing

Preprocessing is conducted sequentially to transform raw comments into structured text ready for analysis. The stages include data cleaning, case folding, normalization, tokenization, stopword removal, and stemming. Table 4 presents an example of the preprocessing results for a representative TikTok comment.

Table 4. Pre-processing Steps (Sample from the TikTok Comments Dataset)

Step	Result
Text Comment	Stop!! kalian sedang melakukan cyber bullying.
Cleaning	Stop kalian sedang melakukan cyber bullying
Case Folding	stop kalian sedang melakukan cyber bullying
Normalization	stop kalian sedang melakukan cyber bullying
Tokenizing	[stop,kalian, sedang, melakukan, cyber,bullying]
Stopword Removal	[stop, cyber, bullying]
Stemming	stop cyber bullying

3.2.3. Word Frequency

This stage analyzes word frequency to determine the urgency and dominance of information in the text. The results serve as a basis for feature selection and data cleaning, removing meaningless words to improve the accuracy and efficiency of the analysis. Figure 2 shows the word frequency before data preprocessing, and Figure 3 shows the word frequency after the data preprocessing stage.

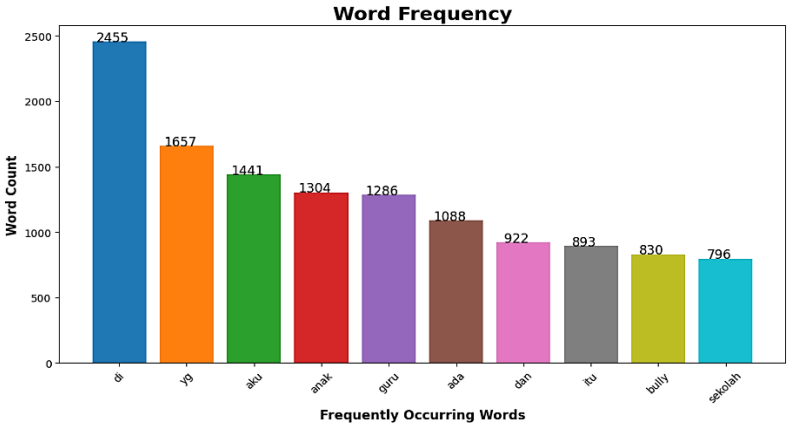

Figure 2. Word Frequency Before Pre-processing

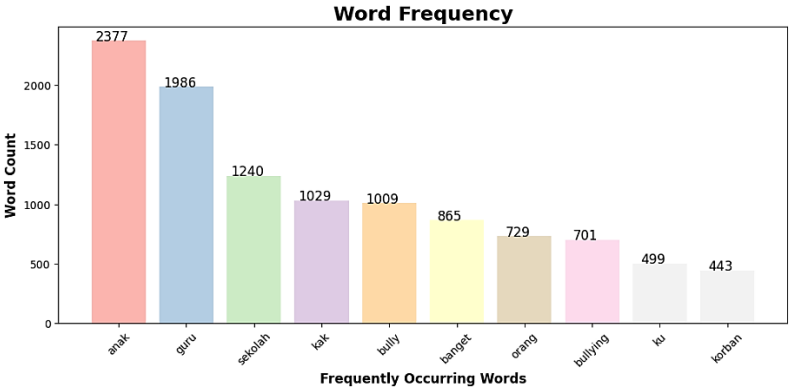

Figure 3. Word Frequency After Pre-processing

3.2.4. Data Labeling

After preprocessing, labeling was performed on 7,900 datasets to be used as training data through the classification of positive, neutral, and negative sentiments.

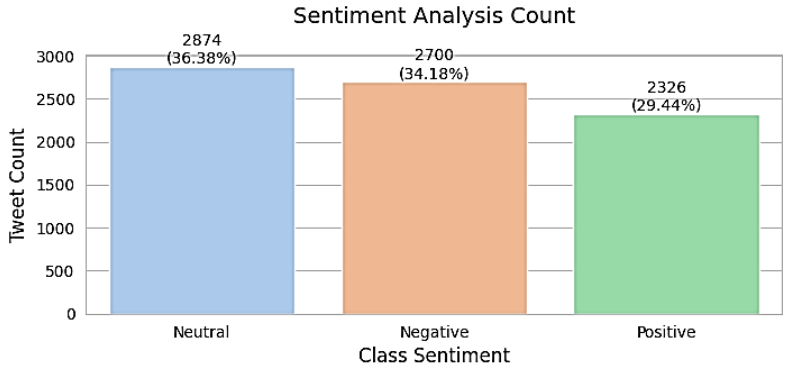

Figure 4. Visualization of Data Labeling Results

The visualization of the labeling results in Figure 4 shows the distribution of the data: 2,874 neutral comments (36.38%), 2,700 negative comments (34.18%), and 2,326 positive comments (29.44%). This distribution indicates that the majority of TikTok users tend to express neutral views regarding the analyzed cyberbullying cases.

3.3. Classification of SVM, KNN, and Naïve Bayes

This research applies SVM, KNN, and Naive Bayes algorithms to classify the sentiment of TikTok comments regarding cyberbullying cases [17], [18]. Naive Bayes uses a probabilistic approach to predict sentiment categories based on word frequency, while SVM and KNN serve as performance comparators

through kernel functions and distance-based proximity between data points. The dataset, sourced from TikTok user comments on cyberbullying cases, is split at an 80:20 ratio, with 80% allocated to training and 20% to testing.

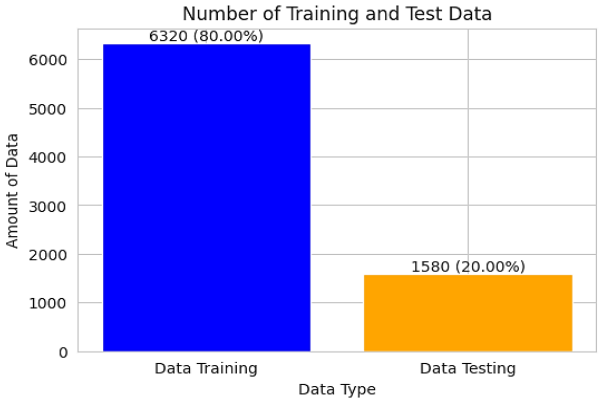

Figure 5. Visualization of The Number of Training And Testing Data

As shown in Figure 5, the total data population comprises 6,320 entries for training and 1,580 for testing to evaluate the prediction accuracy of the three algorithms.

3.4. Evaluation

3.4.1. Confusion Matrix

The classification performance evaluation was conducted using a confusion matrix to assess the model's accuracy. The test results with an 80:20 data ratio for each algorithm are presented in Figure 6, which shows the evaluation matrix for SVM, Figure 7 for KNN, and Figure 8 for Naive Bayes. These results serve as key parameters for assessing each method's classification capability across the available sentiment categories.

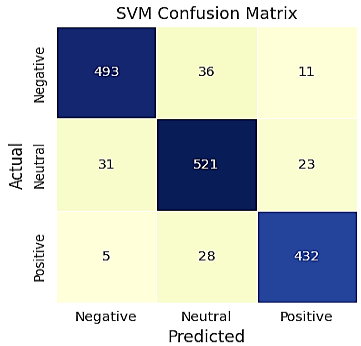

Figure 6. Evaluation Matrix of The SVM Algorithm

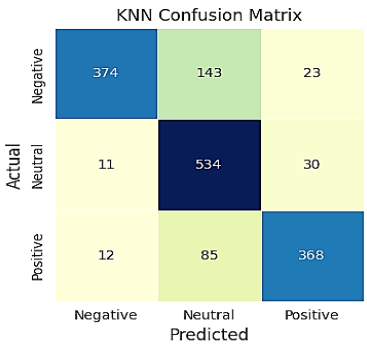

Figure 7. Evaluation Matrix of The KNN Algorithm

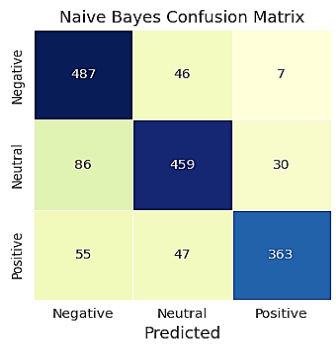

Figure 8. Evaluation Matrix of The Naïve Bayes Algorithm

Based on the test results from these three confusion matrix images, Table 5 summarizes the number of test data that were correctly classified as true positives by each algorithm.

Table 5. Comparison of Correct Predictions on 80:20 Split

Algorithm	Negative (True Positive)	Neutral (True Positive)	Positive (True Positive)	Total Correct Predictions
SVM	493	521	432	1,446
KNN	374	534	368	1,276
Naive Bayes	487	459	363	1,309

In multi-class classification, a confusion matrix illustrates a model's prediction accuracy against actual labels. Key terms include: True Positive (TP), for instances correctly classified as their actual class; True Negative (TN), for instances correctly identified as not belonging to a class; False Positive (FP), for instances incorrectly predicted as a class; and False Negative (FN), for instances of a class incorrectly missed. Table 5 specifically shows the True Positives (diagonal elements) for each sentiment category.

Analyzing Table 9's results for the 1,580 test data, the three algorithms showed varying classification performance. SVM was the most consistent model, making 1,446 correct predictions and excelling at minimizing errors across all sentiment labels. While KNN performed well in identifying neutral sentiment (534 instances), it exhibited a higher error rate in the negative category. Naive Bayes showed stable performance with 1,309 accurate predictions, though its error distribution was broader than SVM's. Consequently, SVM demonstrated the best performance in categorizing cyberbullying sentiments in TikTok comments in this study, surpassing both Naive Bayes and KNN.

3.4.2. Accuracy and Error

Accuracy represents the model's predictive precision, while error indicates the frequency of classification errors. Both metrics are used to evaluate the effectiveness of the SVM, KNN, and Naive Bayes methods in processing data [19]. Figure 9 compares the accuracy and error values of these three algorithms using an 80% training and 20% test split, providing a comprehensive overview of the model's performance.

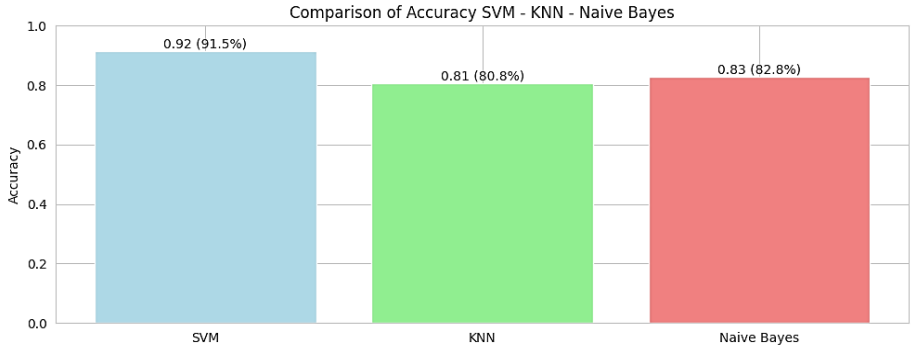

Figure 9. Visualization of the Accuracy Results of SVM, KNN, and Naïve Bayes Algorithms

Based on the comparison accuracy graph you attached, here is a table comparing the Accuracy and Error Rate values for the three algorithms, with Error Rate calculated as 100% - Accuracy.

Table 6. Comparison of Accuracy and Error Rate

Algorithm	Accuracy (%)	Error Rate (%)
SVM	91,5%	8,5%
KNN	80,8%	19,2%
Naive Bayes	82,8 %	17,2%

Table 6 above shows the performance evaluation results of the three algorithms tested. SVM achieved the highest accuracy rate of 91.5% with the lowest error rate of 8.5%. Next, Naive Bayes produced an accuracy of 82.8% with an error of 17.2%. Meanwhile, KNN had the lowest accuracy among the three, at 80.8% with an error rate of 19.2%. These results indicate that SVM is the most optimal model for classifying data.

3.4.3. Classification Report

A detailed analysis of the data classification capabilities was conducted by reviewing the metrics of precision, recall, F1-score, and support for each sentiment label [11]. This classification report maps specific performance and the strengths and weaknesses of the SVM, KNN, and Naive Bayes algorithms. As shown in Figures 10, 11, and 12, the results of this evaluation provide a comprehensive overview of how effectively each model recognizes the characteristics of Positive, Neutral, and Negative sentiments.

	precision	recall	f1-score	support
Negatif	0.932	0.913	0.922	540.000
Netral	0.891	0.906	0.898	575.000
Positif	0.927	0.929	0.928	465.000
accuracy	0.915	0.915	0.915	0.915
macro avg	0.917	0.916	0.916	1580.000
weighted avg	0.915	0.915	0.915	1580.000

Figure 10. Classification Report of The SVM Algorithm

	precision	recall	f1-score	support
Negatif	0.942	0.693	0.798	540.000
Netral	0.701	0.929	0.799	575.000
Positif	0.874	0.791	0.831	465.000
accuracy	0.808	0.808	0.808	0.808
macro avg	0.839	0.804	0.809	1580.000
weighted avg	0.834	0.808	0.808	1580.000

Figure 11. Classification Report of The KNN Algorithm

	precision	recall	f1-score	support
Negatif	0.775	0.902	0.834	540.000
Netral	0.832	0.798	0.815	575.000
Positif	0.907	0.781	0.839	465.000
accuracy	0.828	0.828	0.828	0.828
macro avg	0.838	0.827	0.829	1580.000
weighted avg	0.835	0.828	0.828	1580.000

Figure 12. Classification Report of The Naïve Bayes Algorithm

Based on the Classification Report image for the three algorithms, here is a summary in tabular form. Table 7 presents the weighted average values to provide an overall picture of the model's performance based on the data proportions.

Table 7. Comparison of the Classification Report (Weighted and Macro Average) - 80:20 Split

Algorithm	Metric	Negative	Neutral	Positive	Weighted Avg	Macro Avg
SVM	Precision	0.932	0.891	0.927	0.915	0.917
	Recall	0.913	0.906	0.929	0.915	0.916
	F1-Score	0.922	0.898	0.928	0.915	0.916
KNN	Precision	0.942	0.701	0.874	0.834	0.839
	Recall	0.693	0.929	0.791	0.808	0.804
	F1-Score	0.798	0.799	0.831	0.808	0.809
Naive Bayes	Precision	0.775	0.832	0.907	0.835	0.838
	Recall	0.902	0.798	0.781	0.828	0.827
	F1-Score	0.834	0.815	0.839	0.828	0.829

Based on the evaluation results in Figures 10, 11, 12, and Table 7, SVM demonstrated superior overall performance, consistently achieving 0.915 across weighted average precision, recall, and F1-score. Its high macro-average scores further indicate balanced classification across all sentiment categories. In contrast, KNN generally yielded the lowest results due to its sensitivity to noise, while Naive Bayes, despite reasonable performance, was limited by its strong independence assumptions that are often unsuitable for natural language.

3.4.4. Sentiment Wordcloud

The visualization of keywords from the sentiment analysis is presented through a word cloud, distinguishing between positive, neutral, and negative sentiments. The visual representations are shown in Figure 13 for positive sentiment, Figure 14 for neutral sentiment, and Figure 15 for negative sentiment.

Figure 13. Wordcloud Display for Positive Sentiment

Figure 14. Wordcloud Display for Neutral Sentiment

Figure 15. Wordcloud Display for Negative Sentiment

From the three sentiment word cloud images above, here is a table identifying the keywords from the three WordClouds along with their explanations.

Table 8. Dominant Keywords Based on Sentiment Category

Category	Dominant Keywords	Characteristics
Positive	Bahagia, Sukses, Semangat, Senang, Terbaik, Positif, Teman	Contains words of support, appreciation, and positive emotions.
Neutral	Biasa, Guru, Sekolah, Anak, Peduli, Tenang, Tidak	Contains common nouns or neutral statements.
Negative	Membenci, Marah, Benci, Kesal, Salah, Sakit, Kesalahan, Bully	Contains curse words, expressions of anger, and dislike.

The visualizations clearly map the most frequently occurring words by sentiment. Positive sentiment (green) highlights terms such as 'bahagia' (happy), 'sukses' (success), and 'semangat' (spirit), indicative of constructive interactions. Neutral sentiment (blue) is characterized by general and objective words like 'biasa' (normal), 'guru' (teacher), and 'sekolah' (school), which do not carry specific emotional loads. In stark contrast, negative sentiment (red) is dominated by vocabulary expressing anger and hatred, such as 'membenci' (to hate), 'marah' (angry), and 'kesal' (annoyed) [20], which serve as strong indicators of cyberbullying actions. Through the differences in color and word size, it is clear that each type of sentiment has a very distinct communication character [21].

3.5. Discussion and Analysis

A comparative analysis of the SVM, Naive Bayes, and KNN algorithms reveals notable differences in performance for detecting cyberbullying sentiments on TikTok. Referring to the classification report in Figures 10, 11, and 12, SVM (Support Vector Machine) achieved the best performance, with an accuracy of 91.5%. Its superiority is reflected in the consistent precision, recall, and F1-score values of 0.91 across all sentiment classes. This result indicates that SVM is effective in constructing an optimal hyperplane to separate complex and high-dimensional text data, thereby improving classification accuracy for positive, neutral, and negative comments. These findings are consistent with previous studies reporting that SVM performs well in text classification tasks, especially when dealing with sparse and noisy social media data [22].

In contrast, Naive Bayes achieved an accuracy of 82.8%. The model showed strong sensitivity in detecting negative utterances, with a recall of 0.902, indicating that most bullying-related comments were successfully identified. However, its precision remained lower because the model tended to misclassify neutral or positive comments as negative. This limitation is likely due to the conditional independence assumption of Naive Bayes, which often does not fully capture the informal and context-dependent language commonly found in social media comments. Similar tendencies have also been reported in earlier studies, where Naive Bayes performed reasonably well but remained less robust than margin-based methods for sentiment analysis.

Meanwhile, KNN (K-Nearest Neighbors) produced the lowest performance, with an accuracy of 80.8%. Although KNN achieved very high precision for negative sentiment (0.942), its low recall (0.693)

indicates that many bullying comments were not detected and were instead classified as neutral. This imbalance shows that KNN is more sensitive to noisy data and shifting distribution patterns in TikTok comments. As a result, the model is less suitable for datasets with informal language and uneven class boundaries [23]. Compared with previous studies, this behavior is also consistent with the known weakness of distance-based classifiers in handling sparse and high-dimensional text representations [23], [24].

Overall, these findings indicate that SVM is the most suitable model for automated cyberbullying detection on TikTok. From a practical perspective, SVM can be used as the main engine in content moderation systems to reduce misclassification errors and help identify harmful comments more efficiently. This is particularly important for social media platforms that require fast and reliable moderation to support user safety. However, this study has several limitations. First, the analysis is based on data from only one platform, which may limit generalizability. Second, the study relies on conventional machine learning models, which may not fully capture sarcasm, implicit bullying, or deeper contextual meaning. Therefore, future research should explore Large Language Models (LLMs) using prompt-based classification [25], as well as transformer-based models such as RoBERTa, BERT, and IndoBERT [26], [27]. In addition, multimodal features such as image-based detection [28], [29], enhanced Quantum Long Short-Term Memory (QLSTM) [30], specialized Bi-GRU techniques [31], and bot identification may further improve system performance.

4. CONCLUSION

This research examined 7,900 cleaned data points extracted from 10,188 TikTok comments related to cyberbullying. The results show that neutral sentiment was the most dominant category (36.38%), followed by negative sentiment (34.18%), indicating that TikTok comment interactions frequently contain both neutral responses and harmful verbal expressions. Among the tested models, Support Vector Machine (SVM) produced the best performance with an accuracy of 91.5%, outperforming Naive Bayes (82.8%) and K-Nearest Neighbors (80.8%). This indicates that SVM is more effective at handling the high dimensionality and noise in social media text. The main contributions of this study are quantitative mapping of cyberbullying behavior in TikTok comments, comparisons of classical machine learning methods for Indonesian text classification, and a basis for developing automatic content moderation systems.

This study also has several limitations. First, the analysis is limited to a single platform, so the findings may not fully represent cyberbullying patterns across other social media platforms. Second, relying on classical machine learning models may limit the ability to capture context, sarcasm, and implicit bullying. Therefore, future research should explore transformer-based models such as BERT, IndoBERT, or RoBERTa, which are better suited for understanding contextual meaning in text. Subsequent studies should also consider cross-platform datasets, multimodal approaches combining text and image analysis, and bot-detection techniques to improve classification robustness. In addition, integrating explainable AI methods could make model decisions more transparent and useful for moderation systems. Overall, this study provides a foundation for more accurate and scalable cyberbullying detection in digital environments.

REFERENCES

[1] A. Alamsyah and Y. Sagama, "Empowering Indonesian internet users: An approach to counter online toxicity and enhance digital well-being," Intelligent Systems with Applications, vol. 22, p. 200394, Jun. 2024, doi: 10.1016/j.iswa.2024.200394.

[2] A. Mishra, S. Sinha, and C. P. George, "Shielding against online harm: A survey on text analysis to prevent cyberbullying," Eng. Appl. Artif. Intell., vol. 133, p. 108241, Jul. 2024, doi: 10.1016/j.engappai.2024.108241.

[3] H. Yunhao, E. Sophie, C. Elizabeth M., and K. Bianca, "Player versus Player: A systematic review of cyberbullying in multiplayer online games," Computers in Human Behavior Reports, vol. 18, p. 100675, May 2025, doi: 10.1016/j.chbr.2025.100675.

[4] Z. Dong, Z. Wu, and X. Sun, "Follow the herd or your heart? The role of trait mindfulness in adolescents' responses to observed cyberbullying," Pers. Individ. Dif., vol. 243, p. 113228, Sep. 2025, doi: 10.1016/j.paid.2025.113228.

[5] T. Mahmud, M. Ptaszynski, J. Eronen, and F. Masui, "Cyberbullying detection for low-resource languages and dialects: Review of the state of the art," Inf. Process. Manag., vol. 60, no. 5, p. 103454, Sep. 2023, doi: 10.1016/j.ipm.2023.103454.

[6] T. H. Teng, K. D. Varathan, and F. Crestani, "A comprehensive review of cyberbullying-related content classification in online social media," Expert Syst. Appl., vol. 244, p. 122644, Jun. 2024, doi: 10.1016/j.eswa.2023.122644.

[7] Y. Y. Zandroto, A. V. Vitianingsih, A. L. Maukar, N. K. Hikmawati, and R. Hamidan, "Sentiment Analysis of BCA Mobile App Reviews Using K-Nearest Neighbor and Support Vector Machine Algorithm," Indonesian Journal of Artificial Intelligence and Data Mining, vol. 8, no. 2, p. 448, Aug. 2025, doi: 10.24014/ijaidm.v8i2.37773.

[8] O. S. Jelni, M. L. Radhitya, G. W. Wardhana, Ni Wayan Jeri Kusuma, and N. M. M. R. Desmayani, "Sentiment Analysis of BRImo Reviews on Google Play Store Using SVM and KNN," Indonesian Journal of Data and Science, vol. 6, no. 3, pp. 548–562, Dec. 2025, doi: 10.56705/ijodas.v6i3.365.

[9] Ni Wayan Indah Juliandewi, A. S. Kusuma, K. M. D. Putri, I. G. A. Indrawan, and I. G. A. A. M. Aristamy, "Comparison of Naïve Bayes and Random Forest in Sentiment Analysis of State-Owned Banks Management by Danantara on X and YouTubeComparison of Naïve Bayes and Random Forest in Sentiment Analysis of State-Owned Banks Management by Danantara on X and YouTube," Indonesian Journal of Data and Science, vol. 6, no. 3, pp. 527–537, Dec. 2025, doi: 10.56705/ijodas.v6i3.366.

[10] J. Fillies et al., "A novel German TikTok hate speech dataset: far-right comments against politicians, women, and others," Discover Data, vol. 3, no. 1, p. 4, Mar. 2025, doi: 10.1007/s44248-025-00020-y.

[11] M. Alzaqebah et al., "Cyberbullying detection framework for short and imbalanced Arabic datasets," Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 8, p. 101652, Sep. 2023, doi: 10.1016/j.jksuci.2023.101652.

[12] A. Akhter, U. K. Acharjee, Md. A. Talukder, Md. M. Islam, and M. A. Uddin, "A robust hybrid machine learning model for Bengali cyber bullying detection in social media," Natural Language Processing Journal, vol. 4, p. 100027, Sep. 2023, doi: 10.1016/j.nlp.2023.100027.

[13] A. M. Alduailaj and A. Belghith, "Detecting Arabic Cyberbullying Tweets Using Machine Learning," Mach. Learn. Knowl. Extr., vol. 5, no. 1, pp. 29–42, Jan. 2023, doi: 10.3390/make5010003.

[14] N. I. Boyko and V. Yu. Mykhailyshyn, "K-Nn's Nearest Neighbors Method For Classifying Text Documents By Their Topics," Radio Electronics, Computer Science, no. 3, p. 83, Oct. 2023, doi: 10.15588/1607-3274-2023-3-9.

[15] B. Satya, M. H. S J, M. Rahardi, and F. F. Abdulloh, "Sentiment Analysis of Review Sestyc Using Support Vector Machine, Naive Bayes, and Logistic Regression Algorithm," in 2022 5th International Conference on Information and Communications Technology (ICOIACT), IEEE, Aug. 2022, pp. 188–193. doi: 10.1109/ICOIACT55506.2022.9972046.

[16] S. Tuarob, M. Satravisut, P. Sangtunchai, S. Nunthavanich, and T. Noraset, "FALCoN: Detecting and classifying abusive language in social networks using context features and unlabeled data," Inf. Process. Manag., vol. 60, no. 4, p. 103381, Jul. 2023, doi: 10.1016/j.ipm.2023.103381.

[17] O. S. Jelni, M. L. Radhitya, G. W. Wardhana, Ni Wayan Jeri Kusuma, and N. M. M. R. Desmayani, "Sentiment Analysis of BRImo Reviews on Google Play Store Using SVM and KNN," Indonesian Journal of Data and Science, vol. 6, no. 3, pp. 548–562, Dec. 2025, doi: 10.56705/ijodas.v6i3.365.

[18] T. H. Teng, K. D. Varathan, and F. Crestani, "A comprehensive review of cyberbullying-related content classification in online social media," Expert Syst. Appl., vol. 244, p. 122644, Jun. 2024, doi: 10.1016/j.eswa.2023.122644.

[19] R. Rahmaddeni and F. Akbar, "Comparison of Naïve Bayes Algorithm, Support Vector Machine and Decision Tree in Analyzing Public Opinion on COVID-19 Vaccination in Indonesia," Indonesian Journal of Artificial Intelligence and Data Mining, vol. 6, no. 1, p. 8, Apr. 2023, doi: 10.24014/ijaidm.v6i1.19966.

[20] M. Al-Hashedi, L.-K. Soon, H.-N. Goh, A. H. L. Lim, and E.-G. Siew, "Cyberbullying Detection Based on Emotion," IEEE Access, vol. 11, pp. 53907–53918, 2023, doi: 10.1109/ACCESS.2023.3280556.

[21] A. Almomani, K. Nahar, M. Alauthman, M. A. Al-Betar, Q. Yaseen, and B. B. Gupta, "Image cyberbullying detection and recognition using transfer deep machine learning," International Journal of Cognitive Computing in Engineering, vol. 5, pp. 14–26, 2024, doi: 10.1016/j.ijcce.2023.11.002.

[22] S. Hu, W. Lei, H. Zhu, and C. Hsu, "Cyberbullying perpetration on social media: A situational action perspective," Information & Management, vol. 61, no. 6, p. 104013, Sep. 2024, doi: 10.1016/j.im.2024.104013.

[23] R. Alsheikh, E. Fadel, and N. Akkari, "An Adaptive State Consistency Architecture for Distributed Software-Defined Network Controllers: An Evaluation and Design Consideration," Applied Sciences, vol. 14, no. 6, p. 2627, Mar. 2024, doi: 10.3390/app14062627.

[24] A. C. Roy, T. Mahmud, and T. Abrar, "A multi-class cyberbullying classification on image and text in code-mixed Bangla-English social media content," Natural Language Processing Journal, vol. 13, p. 100191, Dec. 2025, doi: 10.1016/j.nlp.2025.100191.

[25] S. Cirillo, D. Desiato, G. Polese, G. Solimando, V. Sugumaran, and S. Sundaramurthy, "Exploring the ability of emerging large language models to detect cyberbullying in social posts through new prompt-based classification approaches," Inf. Process. Manag., vol. 62, no. 3, p. 104043, May 2025, doi: 10.1016/j.ipm.2024.104043.

[26] A. A. Jamjoom, H. Karamti, M. Umer, S. Alsubai, T.-H. Kim, and I. Ashraf, "RoBERTaNET: Enhanced RoBERTa Transformer Based Model for Cyberbullying Detection With GloVe Features," IEEE Access, vol. 12, pp. 58950–58959, 2024, doi: 10.1109/ACCESS.2024.3386637.

[27] M. K. Mali et al., "Automatic detection of cyberbullying behaviour on social media using Stacked Bi-Gru attention with BERT model," Expert Syst. Appl., vol. 262, p. 125641, Mar. 2025, doi: 10.1016/j.eswa.2024.125641.

[28] A. Almomani, K. Nahar, M. Alauthman, M. A. Al-Betar, Q. Yaseen, and B. B. Gupta, "Image cyberbullying detection and recognition using transfer deep machine learning," International Journal of Cognitive Computing in Engineering, vol. 5, pp. 14–26, 2024, doi: 10.1016/j.ijcce.2023.11.002.

[29] S. Ullah, M. Kukreti, A. Sami, M. R. Shaukat, and A. Dangwal, "The role of bystander behavior and employee resilience in mitigating workplace cyberbullying impacts on employee innovative performance," Human Systems Management, vol. 44, no. 4, pp. 629–640, Jul. 2025, doi: 10.1177/01672533251317066.

[30] K. Subhashree and S. M. Kumar, "Enhanced quantum long short-term memory neural network based multi-task learning for sentimental analysis and cyberbullying detection," Expert Syst. Appl., vol. 282, p. 127555, Jul. 2025, doi: 10.1016/j.eswa.2025.127555.

[31] M. Karpagam et al., "An effective cyberbullying-flashing identification on whatsapp using PTS-GReLU-GRU with harmful level prediction," Sci. Rep., vol. 16, no. 1, p. 80, Dec. 2025, doi: 10.1038/s41598-025-28765-1.

BIBLIOGRAPHY OF AUTHORS

Celestina Florecita Mariwy is an undergraduate student at the Department of Informatics Engineering, Faculty of Engineering, University of Papua. Her academic focus lies in informatics skills development and data-driven projects. Throughout her studies, she has been actively involved in various research initiatives and campus organizations, demonstrating a strong commitment to both academic excellence and practical implementation in the field of information technology.

Mrs. Lorna Yertas Baisa, S.T., M.Kom., is a senior lecturer in the Department of Informatics Engineering, Faculty of Engineering, University of Papua. Her academic expertise and research interests primarily focus on Data Mining and advanced informatics subjects. In addition to her teaching responsibilities, she is deeply committed to mentoring undergraduate and graduate students in their research projects. She also plays a pivotal role in curriculum development and is actively involved in securing research grants and fostering strategic academic partnerships to advance the informatics field.

Mr. Andreas Leonardo Sumendap, S.T., M.T., is a lecturer in the Department of Informatics Engineering, Faculty of Engineering, University of Papua, where he also serves as the Head of the Informatics Engineering Laboratory. His pedagogical approach emphasizes practical, hands-on technical training through laboratory-based courses. In addition to managing lab operations and infrastructure, he is actively involved in supervising undergraduate and graduate research projects. His professional contributions extend to collaborative research initiatives and participation in various academic grant activities.