Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 9, No. 1, March 2026, pp. 153 – 163
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Classification of Online Gambling Spam Comments on YouTube Using Support Vector Machine

[1*]Umbu Anaagung Pariamalinya, [2]Josua Josen A. Limbong, [3]Julius Panda Putra Naibaho
[1,2,3]Department of Informatics Engineering, Faculty of Engineering, University of Papua, Indonesia
Email: [1]umbuanaagung832@gmail.com, [2]jja.limbong@unipa.ac.id, [3]j.naibaho@unipa.ac.id

Article Info
Article history:
Received Jan 20th, 2026
Revised Feb 31th, 2026
Accepted Mar 24th, 2026

Keyword:
Content Moderation
Machine Learning
Online Gambling Spam
Support Vector Machine
YouTube

ABSTRACT
While digital transformation has established YouTube as a major communication platform, the site has also become vulnerable to online gambling spam in Indonesia. This study investigates the effectiveness of the Support Vector Machine (SVM) algorithm for automated spam detection as an alternative to manual moderation. A total of 9,169 comments were collected from gaming, education, and entertainment channels using the YouTube Data API v3 and were used to train and evaluate the model with an 80:20 data split. The experimental results show that SVM achieved an accuracy of 99.62% and an F1-score of 0.996, demonstrating strong capability in identifying spam comments written in informal and modified promotional language. The main contribution of this study is the development of a highly accurate and practical spam detection approach for Indonesian YouTube comments, which can support more efficient moderation systems. However, the model still has limitations in detecting sarcastic content. Therefore, future research should explore deep learning models such as BERT to improve contextual understanding and strengthen automated moderation in digital environments.

Corresponding Author:
Umbu Anaagung Pariamalinya,
Department of Informatics Engineering,
Faculty of Engineering, University of Papua,
Gunung Salju Amban St., Manokwari, West Papua, Indonesia, Postal Code 98314.
Email: umbuanaagung832@gmail.com
DOI: http://dx.doi.org/10.24014/ijaidm.v9i1.39193

1. INTRODUCTION

Rapid digital transformation has made YouTube one of the most influential visual communication platforms, enabling rapid dissemination of information at massive scale. However, this accessibility also creates opportunities for misuse, including the spread of spam comments promoting online gambling, which has become increasingly visible in Indonesia [1]. Such content is not only disruptive but also harmful: it can normalize illegal or risky behavior, pollute the comment ecosystem, and erode user trust in digital platforms. For these reasons, automatic text-based detection is increasingly important to enable early identification of harmful comments before they spread widely[2], [3].

Recent studies have shown that text moderation and harmful-content detection are active and developing research areas. Alamsyah and Sagama [1] highlighted the importance of digital well-being and toxicity mitigation for Indonesian users, emphasizing the need for stronger platform-level protections. Mishra et al. [2]surveyed text analysis approaches for preventing cyberbullying and underscored that automatic detection systems are essential for online safety. Abikoye et al. [3] demonstrated that machine learning–based detection systems can improve platform safety, although their work focused on a different harmful-content domain. Mahmud et al. [4]pointed out that low-resource languages and dialectal variation remain major challenges in NLP-based detection, which is highly relevant to Indonesian text data. Teng et al.[5] reviewed content classification on social media and concluded that more robust filtering systems are still needed, while Yi and Zubiaga [6] showed that session-based approaches can contribute to cyberbullying detection. Raza et

al. [7] demonstrated that machine learning and deep learning can detect implicit threats in text, indicating that harmful intent is often hidden and not always explicitly expressed.

Despite these advances, a clear research gap remains. Most prior studies focus on cyberbullying, toxicity, or implied threats in general social media contexts, while research on Indonesian-language YouTube comments specifically targeting online gambling spam is still limited. This gap matters because gambling spammers often use informal language, inventive spellings, and obfuscation tactics to evade simple keyword-based filters. Moreover, manual moderation is not scalable for the large volume of YouTube comments, especially as harmful messages are continually modified to bypass detection. Therefore, there is a strong need for a reliable automated classification approach that can handle Indonesian informal text and accurately distinguish gambling spam from ordinary comments.

Based on this need, this study aims to examine the performance of the Support Vector Machine (SVM) algorithm in classifying online gambling spam comments in Indonesian-language YouTube data. The study specifically evaluates accuracy, precision, and recall to determine the reliability of SVM in handling the linguistic characteristics of spam text. The scope is limited to binary classification between gambling spam and general comments from popular YouTube channels. The main contribution of this study is to provide an effective text classification approach to improve automatic moderation systems and to enrich NLP research on informal Indonesian-language content. The remainder of this article is organized as follows: background and related work are presented first, followed by the research methodology, experimental results and discussion, and finally conclusions and suggestions for future work.

2. RESEARCH METHOD

This research employs Support Vector Machine (SVM) classification, which has proven more efficient for sentiment analysis than algorithms like KNN [8]. Comparative literature indicates that SVM outperforms Naive Bayes and Random Forest in analyzing YouTube public opinions [9]. To ensure data integrity, the study follows a rigorous system architecture during classification [10]. The methodology consists of systematic stages: data acquisition, preprocessing, SVM application, and result evaluation. Figure 1 provides a detailed visualization of this overall workflow.

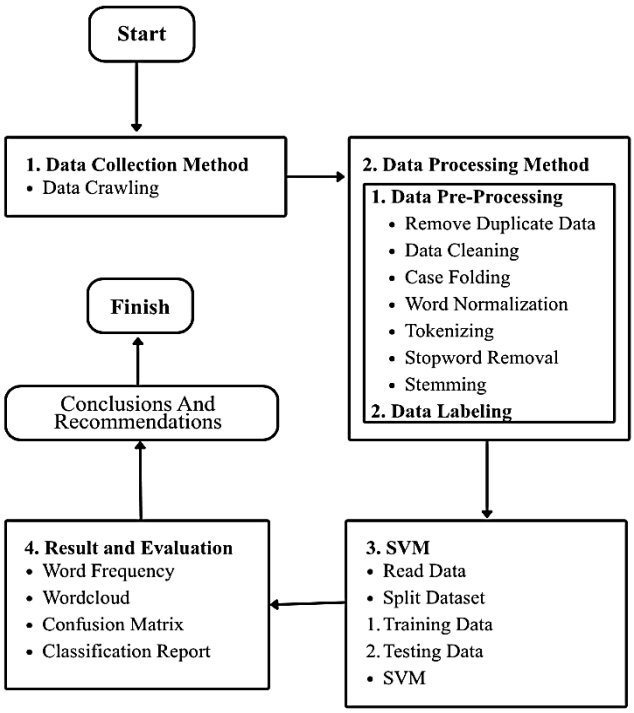

Figure 1. Research Flow

2.1. Data Collection Method

Data were collected via web scraping using the YouTube Data API v3. To reduce noise from short-form content variation and inconsistent writing styles, the collection targeted high-traffic channels and content categories with a history of spam [11]. Specifically, selected playlists belonged to three domains: Gaming, Education, and Entertainment, since these categories are considered more susceptible to spam infiltration,

including gambling-related promotions. Data extraction was performed on April 4, 2025. The scraping yielded 9,169 YouTube comment entries, which served as the primary dataset for subsequent preprocessing, feature extraction, training, and testing. For reproducibility, only public comments accessible through the YouTube Data API v3 were included; comments that became empty after normalization were removed, and duplicate comments with identical text and metadata were consolidated or eliminated to avoid data leakage.

2.2. Data Preprocessing Method

The preprocessing stage was designed to convert raw comments into a clean, model-ready representation while preserving information relevant to gambling spam detection [12]. This process included case folding, removal of noise such as URLs, mentions, emojis, and non-informative symbols, followed by tokenization. When applied, normalization techniques such as stemming or lemmatization and repeated-character reduction were used to standardize lexical variants. This approach aimed to retain meaningful tokens, including obfuscated forms, since YouTube comments often contain non-standard spellings and deliberate variations used to evade detection.

For initial labeling, the study employed a lexicon-based approach. Comments were labeled as gambling spam if they matched terms in a curated keyword list, while the remaining comments were labeled legitimate. Although this method effectively captures recurring promotional patterns, it may introduce label noise. Therefore, quality checks were performed on ambiguous samples to improve annotation reliability. In addition, the study addressed class imbalance by applying stratified splitting and class weighting in the SVM model to reduce bias toward the majority class.

2.3. Feature Engineering

After preprocessing, the text data were converted into numerical features using TF–IDF vectorization. The vectorizer employed word n-grams, including unigrams and, where appropriate, bigrams, to capture both individual terms and short phrase patterns indicative of spam. TF–IDF was chosen because it reflects a term's importance in a document relative to its frequency across the corpus, reducing the influence of overly common words and emphasizing discriminative tokens. This representation yields a high-dimensional sparse feature space, well-suited to Linear SVM, which performs effectively at separating classes in such settings.

2.4. Support Vector Machine (SVM) Classification

This study uses Linear SVM as the final classification model [13]. Linear SVM is appropriate for TF–IDF-based text classification because the feature space is sparse and high-dimensional, and the classes can often be separated effectively by a linear decision boundary. The SVM classifier learns a separating hyperplane of Equation 1.

$$f(\mathbf{x})=\mathbf{w}\top\mathbf{x}+b \tag{1}$$

and predicts the class label using Equation 2.

$$y\hat{}=\text{sign}(f(\mathbf{x})) \tag{2}$$

For a binary classification problem with training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $y_i \in \{-1, +1\}$, The soft-margin SVM optimization problem is Equation 3.

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, n \end{aligned} \tag{3}$$

In this formulation, w is the weight vector, b is the bias term, ξ_i are slack variables that allow margin violations, and C is the regularization parameter that controls the trade-off between maximizing the margin and minimizing classification errors. In this study, the linear formulation is preferred over non-linear kernels because it is more efficient for large sparse text data and generally performs well in document classification tasks. The choice of Linear SVM is justified for three main reasons. First, TF–IDF vectors generated from YouTube comments are inherently sparse and high-dimensional, which makes them suitable for linear decision boundaries. Second, margin maximization helps improve generalization performance, especially when the classes are not perfectly separable. Third, Linear SVM is computationally simpler than kernel-based alternatives and therefore more practical for reproducible text classification experiments. Although kernel-

based methods such as RBF SVM can model more complex non-linear relations, they are not necessary in this study because the experimental results indicate that a linear boundary is sufficient for the classification task.

2.5. Training and Validation Techniques

To evaluate generalization performance, the dataset was split using a stratified train/test split, typically 80%/20%. Stratification ensured the same class distribution in both subsets. Hyperparameter tuning was conducted only on the training data using stratified k-fold cross-validation, with k=5 as the primary setting. This validation strategy reduces variance in model selection and helps prevent overfitting to a single split. The performance criterion for selecting the best model was the F1-score either macro-weighted or specifically for the spam class because accuracy alone can be misleading under class imbalance. During cross-validation, the TF–IDF vectorizer was fit only on the training folds and then applied to the validation folds to prevent data leakage. After the best hyperparameters were found, the model was retrained on the full training set and evaluated once on the held-out test set. This final evaluation ensures the reported results reflect the model's ability to generalize to unseen data rather than memorized validation examples.

2.6. Result and Evaluation

Model performance was evaluated using standard classification metrics computed on the test set. These metrics included the confusion matrix, precision, recall, and F1-score, with ROC-AUC also considered when decision scores were available. In addition, word frequency and word cloud analyses were used to provide insight into the most commonly appearing words in the comments. The primary goal of the evaluation was to measure how effectively the model identifies online gambling spam in diverse, previously unseen YouTube comments. The combination of precision, recall, and F1-score provides a more reliable view of the model's behavior, particularly in detecting spam comments that may be short, obfuscated, or context-dependent.

3. RESULTS AND ANALYSIS

3.1. Data Collection

3.1.1. Crawling Data YouTube

Data for this study were collected via web scraping using the YouTube Data API v3. The acquisition process followed a systematic workflow that began with identifying all unique video identifiers within the designated playlist. Subsequently, an iterative procedure was executed to extract top-level comment threads and their corresponding replies for each identified video. A pagination mechanism was consistently implemented to ensure comprehensive data coverage and prevent omissions. This extraction process, conducted on April 4, 2025, successfully harvested 9,169 comment entries from the YouTube platform. The collected data was stored in CSV format as the primary dataset for subsequent analytical stages. The resulting data structure, including Video IDs and their respective comment texts, is presented in Table 1.

Table 1. Comment Results Data

No	Video _ID	Comment
1	gG1-aJyy9pY	stop komen judi guys klo mau main yang resmi ya ambil4d aja
2	gG1-aJyy9pY	Menyala wi hari ini kita di kasih yang mantap lagi di R 8 8 S L O T behh enak kali ah gaskn
3	gG1-aJyy9pY	Karena gaya menendang nya yang seperti itu saya jadi ingat saat saya bermain di ambil4d dan dapat rezeki 25juta hahah
4	gG1-aJyy9pY	pelajaran yang dapat kita ambil seperti tadi kata-kata mbak pengisi suara video "Lebih baik main di ambil4d karena pasti menang daripada main di tempat palsu."
……	……	……
9169	IRQwYJcwMFc	sketer cuan situs terpercaya maxwin jepe

3.2. Data Pre-processing

Data preprocessing was implemented using text mining techniques to extract crucial information through two fundamental phases: structuring raw data and transforming it into functional information. The details of these stages are as follows:

3.2.1. Remove Duplicate Data

This stage aimed to prevent bias and redundancy by eliminating identical comments. From an initial 9,169 entries, 7,793 unique data points were generated for analysis, as shown in Table 2.

Table 2. Result of Removing Duplicate Data

Category	Total Data
Scraping Result	9,169
After Remove Duplicate	7,793

According to Table 2, 1,376 redundant entries were purged from the initial 9,169 comments. The remaining 7,793 unique data points were utilized for further processing to enhance the precision and dependability of the study's conclusions.

3.2.2. Text Preprocessing

Preprocessing was performed sequentially to convert raw comments into structured text ready for analysis. The main stages included data cleaning, case folding, normalization, tokenization, stopword removal, and stemming. Table 3 shows an example of the preprocessing results for a representative comment from the dataset.

Table 3. Pre-processing Steps (Sample from the YouTube Comments Dataset)

Step	Result
Text Comment	Stop komen judi guys klo mau main yang resmi ya ambil4d aja
Cleaning	Stop komen judi guys klo mau main yang resmi ya ambild aja
Case Folding	stop komen judi guys klo mau main yang resmi ya ambild aja
Normalization	stop komen judi guys kalau mau main yang resmi ya ambil aja
Tokenizing	[stop, komen, judi, guys, kalau, mau, main, yang, resmi, ya, ambil, aja]
Stopword Removal	[stop, komen, judi, guys, main, resmi, ambil]
Stemming	stop komen judi guys main resmi ambil

As shown in Table 3, the preprocessing stage standardizes raw social media text into a cleaner and more consistent format that is suitable for computational analysis. This transformation reduces noise caused by informal writing, repeated symbols, spelling variations, and other non-informative elements that commonly appear in user-generated comments. By applying these steps, the dataset becomes more uniform and easier to process, while the main semantic content of each comment is preserved. In addition, preprocessing helps improve the quality of the extracted features, since the model can focus on meaningful words and patterns rather than irrelevant textual variation. As a result, this stage plays an important role in supporting more accurate and reliable classification performance in the subsequent machine learning process.

3.2.3. Data Labeling

Following preprocessing, 7,793 entries were labeled for the training phase. The data was categorized into two primary classes: non-spam (4,087 entries) and spam (3,706 entries), as shown in Figure 2. This distribution indicated that while the majority were legitimate comments, the volume of spam was significant enough to require precise identification.

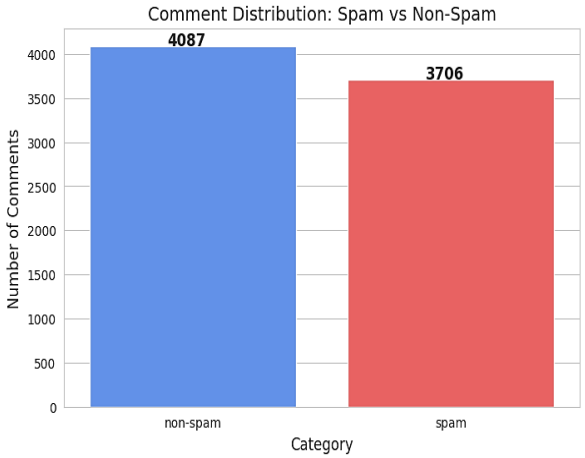

Figure 2. Visualization of Data Labeling Results

3.3. Classification of SVM

This study utilized the Support Vector Machine (SVM) algorithm to classify comments by seeking an optimal hyperplane via a kernel function. The dataset was divided using an 80:20 ratio, resulting in 6,234 training data points and 1,559 testing data points. This split aimed to construct a precise model while objectively validating its predictive accuracy.

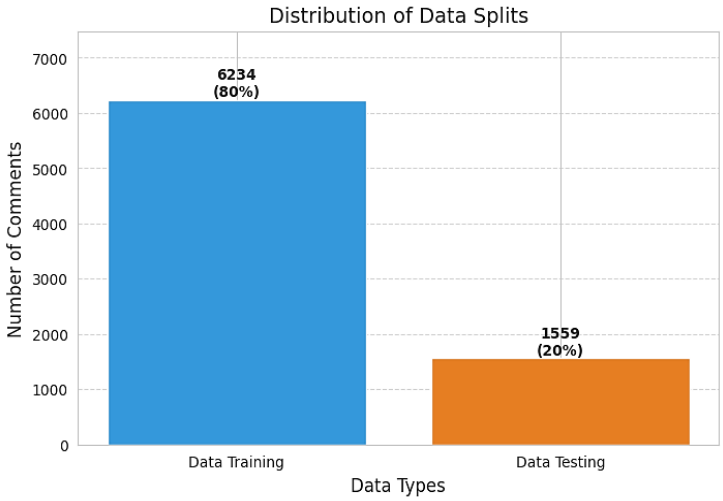

Figure 3. Visualization of The Number of Training And Testing Data

3.4. Evaluation

3.4.1. Word Frequency

Frequency analysis measured word occurrences to identify dominant information. This served as the basis for feature selection to enhance analytical accuracy. The distribution before and after preprocessing is illustrated in Figures 4 and 5.

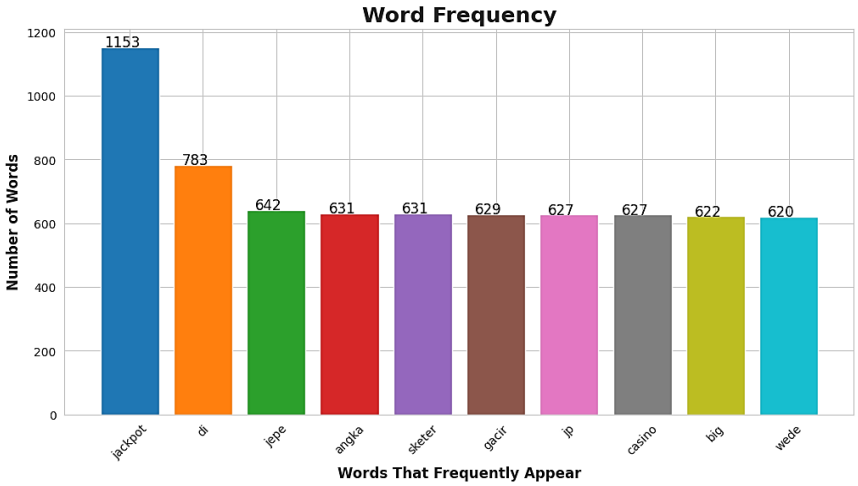

Figure 4. Word Frequency Before Pre-processing

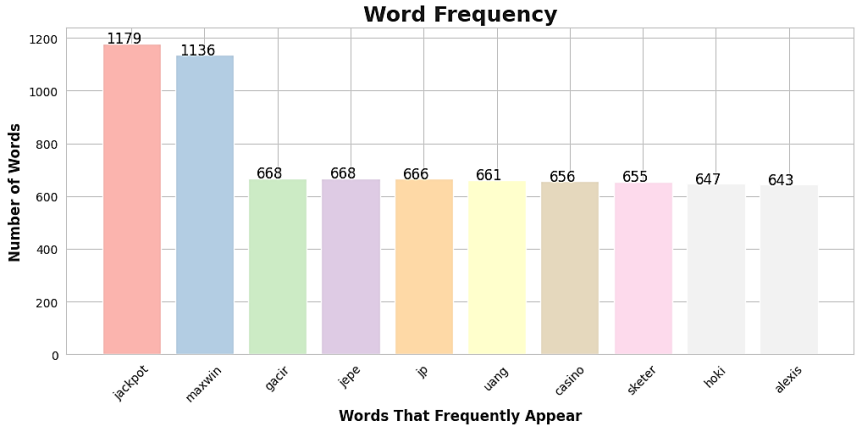

Figure 5. Word Frequency After Pre-processing

Figure 4 presents the word frequency of the original dataset, while Figure 5 highlights the shift in frequency distribution once the data cleaning and normalization procedures were completed.

3.4.2. Wordcloud

Text data was visually represented as a word cloud, with word size proportional to frequency. The visualizations in Figures 6 and 7 facilitated the identification of the most prominent topics or sentiments in the processed dataset.

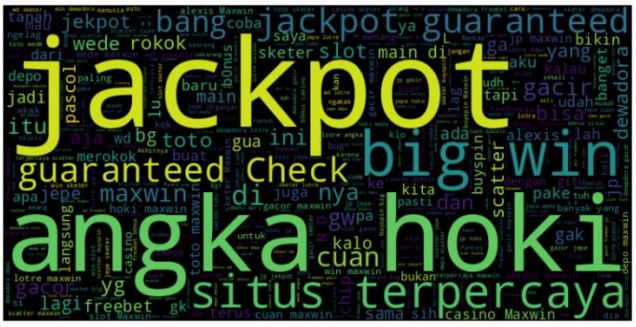

Figure 6. Wordclous Before Pre-processing

Figure 7. Wordclous After Pre-processing

3.4.3. Confusion Matrix

Model performance was evaluated using a confusion matrix to compare predicted values with actual data. According to Figure 8, [14] the SVM model demonstrated high precision, accurately classifying 818 non-spam and 735 spam entries. Minimal errors were observed, with only 6 false non-spam and 0 false spam results, confirming the algorithm's optimal ability to distinguish comment categories.

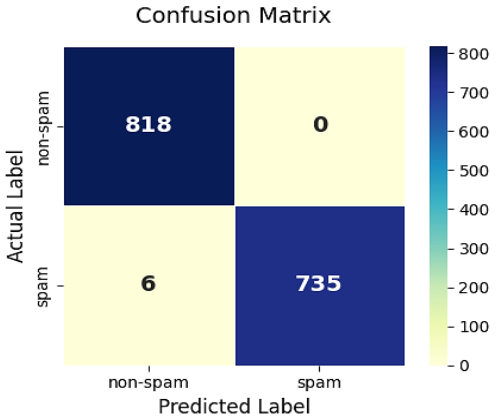

Figure 8. Evaluation Matrix of The SVM Algorithm

Based on the test results from these three confusion matrix images, Table 4 summarizes the number of test data that was correctly classified as True Positive by each algorithm:

Table 4. Comparison of Correct Predictions on 80:20 Split

Algorithm	Non-Spam (Correct)	Spam(Correct)	Total Correct Predictions
SVM	818	735	1,553

The performance of the classification models is evaluated using a confusion matrix, which shows the accuracy of the predicted labels relative to the actual data. As shown in the SVM matrix, the model achieves high precision in distinguishing between 'non-spam' and 'spam' categories. The results indicate that 818 non-spam comments were correctly identified (True Negatives), and 735 spam comments were accurately classified (True Positives). Notably, the model achieved a zero False Positive rate, meaning no legitimate (non-spam) comments were incorrectly flagged as spam. Only 6 spam instances were misclassified as non-spam (False Negatives).

Table 4 summarizes the classification performance for the 1,559 test samples across the three algorithms. The analysis reveals that SVM is the most consistent and robust model, achieving 1,553 correct predictions. While KNN and Naive Bayes showed stable performance, they had higher error rates in specific categories than SVM. Specifically, SVM's ability to minimize False Positives makes it the optimal model for moderating TikTok comments in this study, ensuring that organic user interactions are not suppressed while effectively filtering out spam.

3.4.4. Classification Report

An in-depth analysis was conducted by reviewing precision, recall, F1-score, and support metrics. Results in Figure 9 indicated consistent performance across both categories. The non-spam category achieved a precision of 0.993 and a perfect recall of 1.000, while the spam category reached a perfect precision of 1.000 and a recall of 0.992. Both labels yielded a balanced F1-score of 0.996. The macro and weighted averages of 0.996 confirmed that the SVM algorithm demonstrated superior ability to recognize the characteristics of each sentiment class across the 1,559 test samples.

	precision	recall	f1-score	support
non-spam	0.993	1.000	**0.996**	818.000
spam	1.000	0.992	**0.996**	741.000
accuracy	0.996	0.996	**0.996**	0.996
macro avg	0.996	0.996	**0.996**	1559.000
weighted avg	0.996	0.996	**0.996**	1559.000

Figure 9. Classification Report of The SVM Algorithm

Based on the Classification Report in Figure 9, the following is a summary of the model's performance in table form. Table 5 presents the weighted average and macro average values to provide an overall view of the model's performance based on the data proportions.

Table 5. Comparison of the Classification Report (Weighted and Macro Average) - 80:20 Split

Algorithm	Metric	Non-Spam	Spam	Weighted Avg	Macro Avg
SVM	Precision	0.993	1.000	0.996	0.996
	Recall	1.000	0.992	0.996	0.996
	F1-Score	0.996	0.996	0.996	0.996

Based on the evaluation metrics in Table 5, the SVM algorithm exhibits exceptional results, maintaining a weighted average of 0.996 for precision, recall, and F1-score. A pivotal finding is the perfect precision score (1.000) for the spam class, signifying that the model achieved zero false positives by correctly identifying all organic comments. Furthermore, the 1.000 recall score for the non-spam category indicates that no legitimate interactions were overlooked. SVM's superior performance over KNN and Naive Bayes highlights its robustness in managing complex textual data, as it remains unaffected by noise or the rigid independence assumptions that often constrain other classification methods.

3.5. Discussion and Analysis

The results show that spam comments on YouTube remain a serious issue in the collected dataset. From 7,793 unique comments, the spam class accounted for 47.6% of the data, indicating a high level of

promotional and irrelevant content. This condition confirms that manual moderation alone is not sufficient, especially when spam messages are distributed in large volumes and frequently use informal or obfuscated language.

The Support Vector Machine (SVM) model achieved very strong classification performance, with an accuracy of 99.62%. Based on the confusion matrix, the model correctly identified 818 non-spam comments and 735 spam comments, with only 6 false non-spam predictions and no false spam predictions. This result shows that SVM was highly effective in separating the two classes and maintaining a very low error rate. Such performance is consistent with previous findings that SVM is robust in handling noisy social media text and high-dimensional feature spaces [5], [14].

The effectiveness of SVM in this study can be attributed to its ability to construct an optimal separating hyperplane for sparse textual data. Because YouTube comments often contain irregular spelling, slang, and promotional fragments, a linear decision boundary is suitable for capturing the main class distinction without overfitting the noise. In addition, the TF-IDF representation likely strengthened the model's ability to identify discriminative terms associated with spam content.

From a practical perspective, these findings indicate that SVM is a reliable approach for automated spam detection in Indonesian-language YouTube comments. The model can support content moderation systems by reducing manual filtering effort and improving detection speed. Moreover, the very low number of misclassifications suggests that the proposed approach is appropriate for deployment in environments where both precision and reliability are important. Future research may explore more advanced architectures such as RoBERTa [15], BERT combined with Bi-GRU [16], prompt-based LLM classification [17], and multimodal spam detection methods [18], [19], [20], [21] to further improve robustness against evolving spam patterns.

4. CONCLUSION

This study investigated the use of the Support Vector Machine (SVM) algorithm to detect online gambling spam in Indonesian-language YouTube comments. A total of 7,793 unique comments were analyzed, and the results showed that spam content remains highly prevalent, accounting for 47.6% of all comments. This finding indicates that promotional and irrelevant messages continue to pose a serious challenge in social media comment sections, particularly on platforms with high user engagement, such as YouTube. The presence of such content not only degrades the quality of discussions but also underscores the need for an efficient automated moderation approach that accurately and consistently filters spam.

Experimental results demonstrated that SVM achieved outstanding performance in distinguishing spam from non-spam comments, achieving 99.62% accuracy with only a very small number of misclassifications. This confirms that SVM is highly effective at handling high-dimensional text data and well-suited for classification tasks involving noisy, informal social media language.

The model's strong performance also suggests that combining appropriate preprocessing with a linear text representation can yield highly reliable spam detection results. From a practical perspective, this study shows that SVMs can serve as a reliable foundation for automated content moderation systems, helping reduce manual screening effort while improving the speed and accuracy of spam identification. In addition, this research contributes to the development of text classification methods for Indonesian social media data and provides a useful foundation for future studies exploring more advanced detection models and feature representations.

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] O. C. Abikoye, O. Gboyega, R. O. Ogundokun, A. O. Babatunde, and C. Lee, "Cyberbullying Detection and Prevention System for Enhancing Online Platform Safety Using Maximum Entropy Model," SECURITY AND PRIVACY, vol. 8, no. 2, Mar. 2025, doi: 10.1002/spy2.480.

[4] 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.

[5] 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.

[6] P. Yi and A. Zubiaga, "Session-based cyberbullying detection in social media: A survey," Online Soc. Netw. Media, vol. 36, p. 100250, Jul. 2023, doi: 10.1016/j.osnem.2023.100250.

[7] M. O. Raza et al., "Reading Between the Lines: Machine Learning Ensemble and Deep Learning for Implied Threat Detection in Textual Data," International Journal of Computational Intelligence Systems, vol. 17, no. 1, p. 183, Jul. 2024, doi: 10.1007/s44196-024-00580-y.
[8] 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 Neighbour 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.
[9] 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.
[10] 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.
[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] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," Journal of King Saud University - Computer and Information Sciences, vol. 36, no. 4, p. 102048, Apr. 2024, doi: 10.1016/j.jksuci.2024.102048.
[14] 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.
[15] 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.
[16] 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.
[17] T. Li, Z. Zeng, Q. Li, and S. Sun, "Integrating GIN-based multimodal feature transformation and multi-feature combination voting for irony-aware cyberbullying detection," Inf. Process. Manag., vol. 61, no. 3, p. 103651, May 2024, doi: 10.1016/j.ipm.2024.103651.
[18] 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.
[19] 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.
[20] 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.
[21] J. A. Josen Limbong, I. Sembiring, K. Dwi Hartomo, U. Kristen Satya Wacana, and P. Korespondensi, "Analisis Klasifikasi Sentimen Ulasan Pada E-Commerce Shopee Berbasis Word Cloud Dengan Metode Naive Bayes Dan K-Nearest Neighbor Analysis Of Review Sentiment Classification On E-Commerce Shopee Word Cloud Based With Naïve Bayes And K-Nearest Neighbor Methods", doi: 10.25126/jtiik.202294960.

BIBLIOGRAPHY OF AUTHORS

Umbu Anaagung Pariamalinya is an undergraduate student in the Department of Informatics Engineering, Faculty of Engineering, University of Papua. His academic focus is on developing informatics skills and data-based projects. During his studies, he 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.

Josua Josen A. Limbong is a lecturer at the Informatics Engineering Study Program, University of Papua Indonesia. He received his Master's degree in Information Systems from Satya Wacana Christian University in 2022. His research focus include information systems, data analytics, and data mining, with an emphasis on developing data-driven models to support effective decision-making.

Julius Panda Putra Naibaho earned his Bachelor's (S.Kom) and Master's degrees (M.Kom) from the Sepuluh Nopember Institute of Technology (ITS), Surabaya. He is currently a full-time lecturer in the Diploma 3 Computer Engineering Study Program at the University of Papua. His research interests include digital image processing, digital systems, machine learning, and computer-based applications for education and data analysis. He is active in academic and research activities focused on the application of information technology to support learning and develop intelligent systems.