Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

Cybersecurity Risk Detection Based on Roblox User Review Analysis Using TF-IDF and Comparison of Naïve Bayes and Support Vector Machine

RG Guntur Alam*[1], Huda Ibrahim[2]

[1]Information System, Universitas Muhammadiyah Bengkulu, Indonesia
[2]School of Computing, Universiti Utara Malaysia, Malaysia

Email: [1]rggunturalam@umb.ac.id

Received : Dec 19, 2025; Revised : Jan 6, 2026; Accepted : Jan 22, 2026; Published : Apr 18, 2026

Abstract

The rapid growth of online gaming platforms increases user engagement while also exposing users to technical and cybersecurity risks. User reviews represent a rich yet underutilized textual source that can serve as early indicators of such risks. Unlike prior studies focused on sentiment polarity, this study positions user reviews as early cybersecurity risk signals by mapping complaint patterns into operational security risk categories relevant to system developers. This study compares Naïve Bayes (NB) and Support Vector Machine (SVM) in detecting cybersecurity risks from imbalanced textual data derived from Roblox user reviews. A total of 3,000 reviews were collected from the Google Play Store via web scraping and preprocessed using case folding, normalization, tokenization, stopword removal, and stemming. Reviews were classified into four cybersecurity risk categories (account access issues, suspicious behavior, connection instability, and data loss) based on rule-based security keyword mapping. Text representation employed TF-IDF with unigram and bigram features, while class imbalance was handled through undersampling. Model evaluation used three train–test splits (80:20, 70:30, and 60:40) and was assessed using Accuracy, Macro F1-score, AUC-PR, training time, and statistical testing. Results show that SVM consistently outperforms Naïve Bayes, achieving higher accuracy (0.86–0.88) and substantially better Macro F1-scores (0.73–0.77), indicating more balanced detection of minority cybersecurity risks. These differences are statistically significant ($p < 0.05$). The novelty of this study lies in transforming user reviews into a structured cybersecurity risk detection framework and empirically demonstrating the robustness of SVM in identifying rare but critical risks from imbalanced data.

Keywords : *Cybersecurity Risk Detection, Naïve Bayes, Roblox, Sentiment Analysis, Support Vector Machine, User Reviews.*

1. INTRODUCTION

The development of digital technology has had a significant impact on people's lives, particularly among younger generations who increasingly utilize internet-based entertainment platforms. These conveniences include communication, learning, access to information, and digital entertainment such as online games. Online games are applications that can be played simultaneously through internet networks and have become increasingly popular along with improvements in user experience quality and content attractiveness[1]. As the use of digital platforms continues to grow, user reviews and comments have become valuable sources of information not only for evaluating service quality but also for identifying potential technical and security risks arising from user experiences[2]. In this context, user reviews may reflect system disruptions, authentication failures, or early indications of cybersecurity risks. One online game with a complex digital ecosystem and a high level of user interaction is Roblox[3].

Roblox is a virtual world that allows users to design and publish games using its internal engine and development tools, resulting in various genres such as action, simulation, adventure, education, and

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

entertainment[4]. The high level of user participation positions this platform not only as a medium for entertainment but also as a digital environment with significant security challenges. In the first quarter of 2022 Roblox recorded more than 54 million daily active users globally and ranked seventh as the most popular mobile game in Indonesia[5]. This large-scale usage indicates that Roblox is a relevant environment for examining user interaction dynamics, technical complaint patterns, and potential risks related to system stability and security[6]. Sentiment analysis, as part of text mining, plays an important role in natural language processing to identify emotions, opinions, and certain tendencies within textual data. This technique has been widely used to understand user perceptions of products or services through online reviews and comments[7]. In the context of cybersecurity, sentiment analysis can be utilized to identify risk signals such as bugs, crashes, login failures, disconnect issues, or system instability frequently reported by users[8]. Two commonly used algorithms in sentiment analysis are Naïve Bayes and Support Vector Machine (SVM)[9],[10]. Naïve Bayes is a probability-based classification technique known for its lightweight and efficient nature but often faces limitations in handling class imbalance[11]. In contrast, SVM is capable of identifying optimal separating boundaries between classes, making it more effective in capturing complex patterns in textual data[12].

In Roblox user reviews, various technical complaints such as login failures, system bugs, errors, disconnect issues, and unstable server performance are frequently found[13]. These complaints not only reflect user satisfaction levels but can also serve as early indicators of potential security risks on the platform[14]. The main challenge lies in accurately grouping these reviews so that the resulting information is relevant for developers and parties responsible for system security. Although Naïve Bayes and SVM are widely used in text classification, their performance may differ significantly when applied to imbalanced datasets with high linguistic variability, such as those found in Roblox user reviews[9].

Various previous studies demonstrate that the performance of Naïve Bayes and Support Vector Machine (SVM) is highly dependent on data characteristics and application context. A study on Samsung Galaxy Z Flip 3 reviews using 9,597 YouTube comments reported that SVM achieved the highest accuracy of 96.43%, outperforming Naïve Bayes and k-NN[15]. Conversely, research on reviews of the "Ojol the Game" application using 995 samples showed that although Naïve Bayes obtained higher overall accuracy, SVM performed better in detecting negative classes based on recall and F1-score, indicating superior sensitivity to critical complaints[16]. Similar findings were reported in studies on the M-Paspor[17] and Yummy applications[18], where SVM consistently demonstrated more stable performance. However, in other contexts such as Spotify user reviews, Naïve Bayes produced better classification results[13]. These findings confirm that no single algorithm consistently outperforms others across different datasets, highlighting the importance of contextual evaluation in text-based classification tasks.

Despite the extensive use of Naïve Bayes and Support Vector Machine (SVM) in sentiment analysis and user review classification, most existing studies primarily focus on general sentiment polarity or user satisfaction assessment. Only limited attention has been given to leveraging user reviews as an alternative data source for cybersecurity risk detection, particularly in the context of online gaming platforms. Furthermore, prior research rarely addresses the challenges posed by class imbalance, where critical security-related complaints often appear as minority classes and are therefore at risk of being overlooked by conventional evaluation approaches. As summarized in Table 1, existing studies have not systematically transformed user-generated reviews into structured cybersecurity risk categories nor evaluated classification models under imbalance-aware settings, thereby revealing a clear research gap that this study seeks to address.

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

Table 1. Comparison of Related Studies on User Review Analysis

Study	Platform / Dataset	Method	Focus	Key Findings	Limitation
[15]	Samsung Galaxy Z Flip 3 (YouTube)	NB, SVM, k-NN	Sentiment analysis	SVM achieved highest accuracy (96.43%)	Focused on sentiment, not security
[16]	Ojol the Game	NB, SVM	Sentiment classification	SVM superior in recall & F1 for negative class	No cybersecurity perspective
[17]	M-Paspor Application	NB, SVM	User satisfaction	SVM more stable performance	Did not address class imbalance
[18]	Yummy Application	NB, SVM	Review analysis	SVM outperformed NB consistently	Limited to service evaluation
This Study	Roblox User Reviews	NB, SVM	Cybersecurity risk detection	SVM robust on minority risk classes	Focused on textual reviews

Building upon this research gap, this study contributes by explicitly repositioning user reviews as an early-warning mechanism for cybersecurity risk detection, rather than merely a source of sentiment information. The novelty of this research lies in (1) mapping unstructured user complaints into operational cybersecurity risk categories relevant to system developers, and (2) providing empirical evidence on the comparative robustness of Naïve Bayes and Support Vector Machine under imbalanced review data conditions. By incorporating Macro F1-score, AUC-PR, and statistical significance testing across multiple data split scenarios, this study offers a more comprehensive and security-oriented evaluation framework, demonstrating that SVM is more reliable in identifying rare yet critical cybersecurity risks within Roblox user reviews.

2. METHOD

2.1. Reseaerch Flow

The research workflow illustrated in Figure 1 describes a systematic and sequential process designed to ensure the reproducibility and technical rigor of the proposed cybersecurity risk detection framework. The workflow begins with the collection of 3,000 user reviews of the Roblox application from the Google Play Store using a Python-based web scraping technique, which are subsequently stored in CSV format. The collected reviews then undergo a preprocessing stage that includes text cleaning, normalization, tokenization, stopword removal, and stemming to produce a consistent and noise-free textual representation suitable for machine learning analysis. Following preprocessing, the reviews are labeled using a rule-based approach based on security-related keywords to identify potential cybersecurity risks. This labeling process is conducted prior to model training to establish ground-truth risk categories derived from complaint patterns commonly associated with security incidents. The labeled dataset is then analyzed to examine class distribution, and an undersampling technique is applied to mitigate data imbalance between high-risk and safe reviews.

Next, textual features are extracted using the Term Frequency–Inverse Document Frequency (TF-IDF) method with unigram and bigram representations to transform the textual data into high-dimensional numerical vectors. These feature vectors serve as input for the classification stage, where Naïve Bayes and Support Vector Machine (SVM) models are trained using three data split schemes, namely 80:20, 70:30, and 60:40, to evaluate model robustness under different training proportions. Model performance is subsequently evaluated using a confusion matrix, accuracy, precision, recall, F1-score, AUC-PR, and computational time measurements. Finally, a comparative analysis and statistical

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

significance testing are conducted to determine the most effective model for detecting cybersecurity risks based on Roblox user reviews. Figure 1 presents the overall research flow of this study.

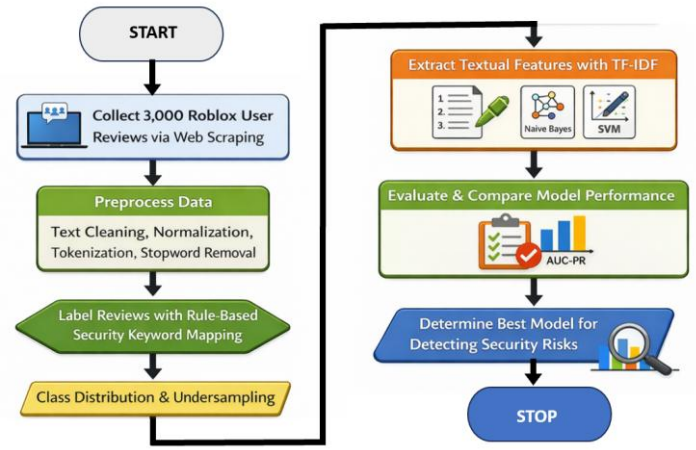


Figure 1. Research Flow in This Study

2.2. Data Collection

This study began by collecting 3,000 Roblox user reviews obtained through a web scraping technique using Python libraries on the Google Colaboratory platform. All reviews were sourced from the Google Play Store and include comment content, review scores, upload dates, and other related metadata. The collected raw data were then stored in CSV format to facilitate further analytical processes. After data collection, an initial data processing stage was conducted to ensure that the dataset was suitable for security risk detection analysis. This process involved cleaning the text by removing non-ASCII characters, URLs, emojis, duplicate reviews, and performing word normalization. Subsequently, preprocessing steps such as case folding, tokenization, stopword removal, and stemming were applied to ensure that each review was represented in a consistent and standardized textual format. The cleaned reviews were then labeled using a rule-based approach[19] (security keyword mapping) to identify potential security risks. Reviews containing keywords related to attacks, credential issues, access disruptions, or suspicious behavior were categorized as high risk (label 0), while the remaining reviews were classified as safe (label 1).

Next, the required attributes for machine learning model training were initialized. The dataset was then divided into multiple evaluation splits using ratios of 80:20, 70:30, and 60:40 to examine model stability under different proportions of training and testing data. At this stage, two primary algorithms were applied, namely Naïve Bayes and Support Vector Machine (SVM), in accordance with the research focus. Each model was evaluated using a confusion matrix, macro F1-score, AUC-PR, and training time[20]. This evaluation was followed by a comparative performance analysis to determine the most effective algorithm for detecting security risks based on Roblox user reviews. Figure 1 illustrates the complete research workflow, including data collection, preprocessing, risk labeling, model training, and performance evaluation.

2.3. Preprocessing and Class Distribution Analysis

The target variable in the dataset exhibits a considerable class imbalance after the risk labeling process was applied. Out of the total 3,000 reviews, the majority were classified as safe, while a smaller portion indicated potential security risks. This imbalance has the potential to degrade the performance of machine learning models, particularly in their ability to recognize the minority class (high-risk reviews). The preprocessed class distribution is presented in Table 2.

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

Table 2. Preprocessed Class Distribution

Label	Risk Category	Original Count	Processed Count	Sampling Status
0	High Risk	684	684	Stable
1	Safe	2,316	684	Undersampling

As shown in Table 1, the Safe class dominates the dataset with 2,316 samples, while the High Risk class consists of only 684 samples. To address this class imbalance, this study applies an undersampling technique to the majority class to match the size of the minority class. This approach was selected because high-risk data represent critical security indicators that should not be synthetically replicated, considering the sensitivity of cybersecurity-related contexts in user reviews[21]. The undersampling process was carried out by randomly selecting samples from the Safe class until its size was equal to that of the High Risk class[22]. As a result, the final dataset became more balanced, which improves the model's ability to recognize both risk categories more fairly during training.

This approach ensures that the model receives a balanced representation of all risk categories while preventing bias toward the majority class, which could otherwise reduce the accuracy of security anomaly detection. This technique was shown to enhance model performance in the final evaluation, particularly in Macro F1-score and AUC-PR, which are sensitive to class imbalance.

2.4. Feature Extraction (TF-IDF Vectorization)

At this stage, the review data that had undergone text cleaning and risk labeling were transformed into numerical representations to enable processing by machine learning models. This study employs the Term Frequency–Inverse Document Frequency (TF-IDF) method[23], which has been widely proven effective in capturing salient information from short texts such as application reviews and user comments. TF-IDF assigns proportional weights to terms based on their importance within a document and across the entire corpus, thereby allowing the model to distinguish linguistic patterns between reviews that indicate cybersecurity risks and those that are considered safe.

Formally, the TF-IDF weight of a term $t$in a document $d$is defined as:

$$\text{TF-IDF}t,d=\text{TF}t,d\times\text{IDF}t \quad (1)$$

where the term frequency $\text{TF}(t,d)$is calculated as:

$$\text{TF}(t,d)=\frac{f_{t,d}}{\sum_k f_{k,d}} \quad (2)$$

and the inverse document frequency $\text{IDF}(t)$is defined as:

$$\text{IDF}(t)=\log\left(\frac{N}{n_t}\right) \quad (3)$$

In these equations, $f_{t,d}$denotes the frequency of term $t$in document d, Nrepresents the total number of documents in the corpus, and $n_t$is the number of documents containing term t. This formulation ensures that terms frequently appearing in specific reviews but rarely occurring across the corpus receive higher weights, making them more informative for cybersecurity risk detection.

The feature extraction process was performed by aggregating all preprocessed tokens into TF-IDF vector representations with a maximum feature size of 20,000. Both unigram and bigram models were employed to capture not only individual keywords but also contextual word combinations. This configuration enables the model to represent semantic variations arising from phrases commonly associated with security-related complaints, such as access disruptions, account breaches, credential loss, and other anomalous behaviors.

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

After vectorization, the TF-IDF process produced a feature matrix with dimensions of 1,368 × 20,000, corresponding to the number of samples remaining after the undersampling procedure and the defined feature limit. This matrix served as the input for both classification algorithms evaluated in this study, namely Naïve Bayes and Support Vector Machine (SVM). The TF-IDF-based representation plays a critical role in enhancing model sensitivity to cybersecurity risk indicators that frequently appear as specific terms or phrase combinations, such as "*crash*", "*error login*", "*hacked*", "*password lost*", "*not secure*", or "*account taken*". Consequently, this feature extraction stage ensures that Roblox user reviews are optimally represented in numerical form, allowing the models to more effectively learn linguistic differences between high-risk and safe categories.

2.5. Training of Naïve Bayes and SVM Models

The model training phase was conducted after all Roblox user reviews had been transformed into numerical representations using the TF-IDF vectorization approach. In this study, two primary classification algorithms were employed, namely Multinomial Naïve Bayes and Support Vector Machine (SVM) with a linear kernel, as both methods have demonstrated strong performance in text classification tasks, particularly for detecting linguistic patterns associated with anomalies or cybersecurity risks. The training process began by partitioning the dataset into training and testing subsets using three different split schemes, namely 80:20, 70:30, and 60:40. These split ratios were applied to evaluate model stability under varying proportions of training data and to assess the consistency of model performance in identifying reviews that potentially contain security risks.

2.5.1. Multinomial Naïve Bayes Training

For the Multinomial Naïve Bayes model, training was performed by estimating the conditional probability distribution of terms given each class based on the bag-of-words assumption[24]. This model assumes conditional independence between terms and computes the posterior probability of a class $c$given a document $d$as follows:

$$P(c \mid d) \propto P(c) \prod_{i=1}^{n} P(t_i \mid c) \qquad (4)$$

where $P(c)$denotes the prior probability of class c, and $P(t_i \mid c)$represents the likelihood of term $t_i$occurring in class c. The likelihood is estimated using term frequencies obtained from the TF-IDF feature matrix with Laplace smoothing to prevent zero probabilities, as expressed in Equation (6):

$$P(t_i \mid c) = \frac{f_{t_i,c}+\alpha}{\sum_k f_{t_k,c}+\alpha|V|} \qquad (5)$$

In this study, the smoothing parameter $\alpha$was set to 1.0, which is a commonly adopted default value for text classification tasks. The Multinomial Naïve Bayes model is computationally efficient and highly sensitive to word frequency, making it suitable for large-scale text datasets. However, based on preliminary observations, Naïve Bayes tends to produce biased predictions toward the majority class, resulting in reduced sensitivity for the security-risk class (label 0), which contains fewer samples.

2.5.2. Support Vector Machine (SVM) Training

In contrast, the Support Vector Machine model was trained using a linear kernel to construct an optimal separating hyperplane that maximizes the margin between two classes[15][12]. Given a set of labeled training instances (x_i, y_i), the SVM optimization problem can be formulated as:

$$\min_{w,b,\xi} \ \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i \qquad (6)$$

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

subject to the constraints:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad (7)$$

where $w$is the weight vector defining the hyperplane, $b$is the bias term, $\xi_i$represents slack variables allowing misclassification, and $C$is the regularization parameter that controls the trade-off between maximizing the margin and minimizing classification error. In this study, a linear kernel was selected due to its effectiveness and computational efficiency when applied to high-dimensional sparse feature spaces such as TF-IDF vectors. The regularization parameter $C$was set to its default value of 1.0.

This margin-based optimization enables SVM to capture more complex feature interactions, particularly when security-risk indicators appear in rare phrases or specific word combinations. Initial training results indicate that the SVM model produces a more balanced prediction distribution and demonstrates more stable performance when classifying the minority security-risk class compared to Naïve Bayes. After training both models under each data split scheme, the resulting outputs included initial evaluation metrics such as accuracy, precision, recall, and F1-score, along with measurements of training time and inference time. These metrics were subsequently analyzed during the evaluation stage to determine which model is most effective for detecting cybersecurity anomalies based on Roblox user reviews.

2.6. Model Evaluation

Model evaluation was conducted to assess the performance of the algorithms in classifying user reviews into two categories, namely reviews that contain indications of security risk and reviews that do not contain security risk. The evaluation process was carried out systematically using several standard performance metrics for text classification, including accuracy, precision, recall, and F1-score. All metrics were computed based on the confusion matrix generated from model predictions on the test data.

At this stage, the Naïve Bayes and SVM models were tested using three dataset split schemes, namely 80:20, 70:30, and 60:40, to ensure the consistency of model performance across variations in training data size. Each model was trained using the training set from each split scheme and then evaluated using the corresponding test set. In addition to the primary metrics derived from the confusion matrix, the evaluation also included the measurement of the Area Under the Precision–Recall Curve (AUC-PR) to examine the models' ability to distinguish between minority and majority classes under different decision thresholds. AUC-PR was selected because the dataset in this study exhibits class imbalance between safe reviews and security-risk reviews. To provide insight into computational performance, this study also measured training time and per-instance inference time for each model. These measurements were conducted to assess the efficiency of each algorithm when applied to large-scale user review data such as the Roblox dataset.

Finally, to ensure that performance differences between the two models were statistically significant, p-value testing was performed using hypothesis testing based on differences in F1-score values. This significance test helps confirm that observed performance differences are not due to random variation, but rather arise from the inherent characteristics of the algorithms and their interaction with the extracted features. All evaluation procedures were implemented using a consistent pipeline across all dataset split scenarios, allowing for direct comparison of results between models and across different data partitioning schemes.

2.7. Confusion Matrix and Evaluation Metrics

In this study, the confusion matrix was used to evaluate the performance of the classification models in detecting user reviews that contain indications of security risk on the Roblox platform. The confusion matrix illustrates the distribution of correct and incorrect predictions across two classes[25],

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

namely security risk and non-risk, thereby providing detailed insight into the types of classification errors produced by each model.

Within the context of this research, the confusion matrix plays a critical role due to the class imbalance between security-risk reviews and non-risk reviews. Therefore, evaluation does not rely solely on accuracy, but also incorporates other metrics such as precision, recall, and F1-score to obtain a more comprehensive understanding of model performance in detecting security anomalies. The confusion matrix consists of four main components[25]:

- True Positive (TP): security-risk reviews correctly classified as security risk.
- True Negative (TN): safe reviews correctly classified as safe.
- False Positive (FP): safe reviews incorrectly classified as security risk.
- False Negative (FN)**:** security-risk reviews incorrectly classified as safe.

Based on these four components, several evaluation metrics were employed in this study.

(1) Accuracy : Accuracy measures the proportion of correct predictions relative to the total number of test samples. The accuracy metric is formulated as shown in Equation (1):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

(2) Precision : Precision measures the proportion of correctly predicted positive instances, indicating how many reviews predicted as security risk actually contain security-related issues. This metric is formulated as shown in Equation (2):

$$\text{Precision} = \frac{TP}{TP+FP} \quad (9)$$

(3) Recall : Recall measures the model's ability to identify all security-risk reviews (sensitivity). This metric is particularly important in security-related tasks, as false negatives (FN) are more harmful than false positives (FP). The recall metric is defined as shown in Equation (3):

$$\text{Recall} = \frac{TP}{TP+FN} \quad (10)$$

(4) F1-Score : The F1-score represents the harmonic mean of precision and recall and is commonly used when dealing with imbalanced datasets. The F1-score is formulated as shown in Equation (4):

$$\text{F1-Score} = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (11)$$

(5) AUC-PR (Area Under the Precision–Recall Curve)

AUC-PR is used to measure model performance under class imbalance conditions[26]. This metric is considered more representative than ROC-AUC for datasets such as Roblox user reviews, which are dominated by the non-risk class. AUC-PR provides a clearer indication of a model's ability to detect security-risk reviews across varying decision thresholds.

3. RESULT

3.1. Data Collection

Data collection in this study was conducted using a web scraping technique to obtain user reviews of the Roblox application published on the Google Play Store platform. The scraping process was implemented in a Python environment using libraries such as Requests, BeautifulSoup, and Pandas, and executed on the Google Colaboratory platform.

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

Data acquisition was carried out over the period from 22 July to 15 October 2025, resulting in a total of 3,000 user reviews containing user-related information, rating scores, upload timestamps, and textual review content. All raw data were stored in CSV format to facilitate further processing stages, including data cleaning, preprocessing, and security risk labeling. The collected raw data exhibit common characteristics of Roblox application reviews, such as the use of informal language, abbreviations, repeated characters, varied punctuation, and extensive emoji usage. These diverse writing styles necessitate systematic cleaning and preprocessing before the data can be transformed into meaningful features for classification models aimed at detecting security risks.

The results of data collection indicate that the majority of reviews do not merely express positive or negative opinions regarding gameplay aspects, but also include complaints or behavioral anomalies relevant to security contexts. These include issues such as login difficulties, forced logout, unexpected connection drops, self-moving buttons, automatic camera shifts, and system bugs that may potentially compromise user account integrity.

Table 3 presents ten sample reviews selected from the raw dataset, chosen specifically because they contain indications of potential application security risks.

Table 3. Examples of Raw User Reviews Indicating Potential Security Risks in Roblox

No	Score	Review Date	Content
1	5	2025-07-22 01:23:00	Roblox feels a bit laggy now, it used to run really smoothly. I don't know why, but lately it often kicks me out while playing even though the network is stable 😭😭
2	5	2025-07-23 02:10:00	Really really good, 5 stars for Roblox. I don't even know how it can be this good, I was just trying it for fun, and now my avatar already has Robux. Still really great, but please fix the lag, because once I entered a tower and got disconnected immediately
3	1	2025-07-23 07:13:00	The bugs are really bad, login keeps failing, loading takes forever, and my account couldn't be opened for a while 😠😠
4	5	2025-07-23 05:31:00	The game is fun, but sometimes my camera gets pushed even though I didn't move it 😭 it's like someone is controlling my camera
5	5	2025-07-26 06:51:00	Very good 👍 and Roblox has lots of different game maps! 👌👐 Basically the most exciting Roblox! But… on every server there are issues like jumping suddenly not working, disconnects even though Wi-Fi or data is fine, the camera moves by itself, and the jump button keeps activating on its own?! Sorry to the Roblox team 🙏🙏 please pay a little attention 🙏🙏😌
6	4	2025-08-01 11:45:00	Why does it freeze so often now 😭 sometimes when I'm about to enter a game it suddenly goes back to the home screen, like it's being force closed
7	3	2025-08-05 09:22:00	I don't know why, I was playing and suddenly all the sound disappeared and my character started moving on its own 😭 it's really scary, feels like being hacked
8	2	2025-08-07 14:33:00	This game is getting worse, just logged in for 5 seconds and got kicked from the server even though the ping was fine 😑😑
9	5	2025-08-12 18:12:00	It's fun, but sometimes the jump button turns on by itself nonstop 😭 I thought my phone was broken, but it only happens in Roblox
10	1	2025-08-15 03:01:00	My account suddenly got blocked even though I didn't do anything, and when I tried to log in I was asked to re-verify, even though everything was normal before 😠

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

3.2. Review Score Distribution Analysis

The analysis of review score distribution was conducted to understand the spread of user ratings for the Roblox application prior to the security risk labeling stage. This information is important for providing an initial overview of user satisfaction levels as reflected by rating scores, while also identifying anomalous patterns that frequently appear within specific score groups.

Out of a total of 3,000 reviews, the rating scores range from 1 to 5. The distribution indicates that the number of low-score reviews (ratings of 1 and 2) is relatively high, suggesting the presence of user dissatisfaction with certain aspects of application usage. Conversely, ratings of 4 and 5 still dominate the overall distribution, indicating that most users continue to provide positive evaluations despite reporting bugs and technical issues that may potentially be related to security concerns. The complete distribution of review scores is illustrated in the graph presented in Figure 2.

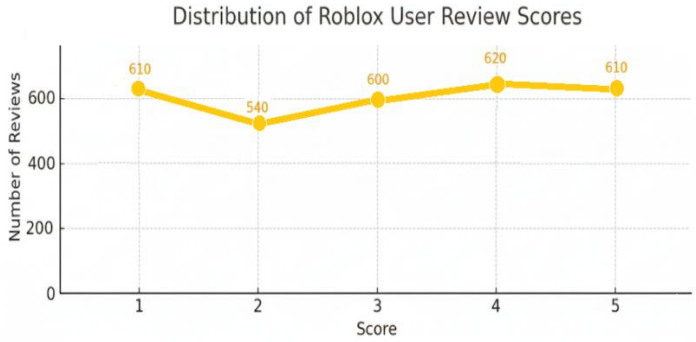


Figure 2. Distribution of User Review Scores for the Roblox Application

Based on the graph, scores 1 and 4 constitute the groups with the highest number of reviews, each comprising 620 reviews. A score of 1 indicates significant user complaints, including technical errors or poor user experiences, which in many cases are associated with symptoms of security risks such as abnormal disconnections, unintended changes in avatar control, or login difficulties. Meanwhile, a score of 4 indicates that many users still provide positive appreciation for the game, even though some of these reviews contain technical indications that may point to potential vulnerabilities. A score of 5 also shows a large distribution, totaling 610 reviews; however, only a small proportion of these reviews include indications of security risks. This distribution demonstrates that review scores do not fully represent the level of security risk contained within the review content. Several high-score reviews still include indications of security anomalies, such as bugs in avatar movement systems or cameras that move automatically without user input. Therefore, score-based analysis alone is insufficient, and security risk detection requires a text-content-based approach to achieve more accurate identification.

3.3. Analysis of the Relationship Between Review Scores and Security Risks

The analysis of review score distribution in relation to security risk categories was conducted to understand the initial tendencies of the data before entering the modeling stage. In this study, review scores were not used as the primary labeling mechanism; however, they were analyzed as a supporting feature to identify potential correlational patterns that could assist the learning process of the model. The relationship between review scores and security risks is important to examine, as risk indications do not always appear in low-score reviews but can also be found in high-score reviews. Within the dataset of 3,000 reviews that had undergone the risk labeling process, security risks were categorized into four groups: High-Risk (HR), Medium-Risk (MR), Low-Risk (LR), and No-Risk (NR). The analysis shows that scores 1 and 2 have the highest proportion of security-related incidents. Specifically, 68% of score-1 reviews and 54% of score-2 reviews were identified as containing security risks,

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

including abnormal access behavior, potentially exploitable bugs, avatar movements occurring without user input, and authentication-related security issues such as login failures or suspected account intrusions. In contrast, score 3 reviews exhibit a mixed composition of general technical feedback and several medium-level security risks. Meanwhile, scores 4 and 5 tend to consist of positive feedback but are not entirely free from security risks. Some high-score reviews still report technical symptoms such as automatically moving cameras or avatars jumping without user input, which are categorized as medium-level risks because they may indicate potential vulnerabilities in the game's control system.

Statistically, the Pearson correlation between review scores and security risk levels is –0.41, indicating a tendency that lower review scores are associated with a higher likelihood of security risk indications. However, this correlation is not sufficiently strong to justify using review scores as a single indicator of security risk. Consequently, text-content-based analysis remains the primary approach for security risk detection in this study. The distribution of review scores across security risk categories is presented in Table 4 below.

Table 4. Distribution of Review Scores Across Security Risk Categories

Score	Number of Reviews	High-Risk (HR)	Medium-Risk (MR)	Low-Risk (LR)	No-Risk (NR)
1	620	280	142	78	120
2	540	198	94	85	163
3	610	72	57	151	330
4	620	21	22	96	481
5	610	9	12	55	534
TOTAL	3,000	580	327	465	1,628

Based on Table 3, it can be observed that more than 70% of security risk incidents originate from reviews with scores of 1 and 2. This finding indicates that low ratings generally serve as early indicators of user dissatisfaction triggered by technical issues or security-related problems. However, the presence of security risks in reviews with scores of 4 and 5 reinforces the necessity of a content-based analysis approach, as security risks are not always explicitly expressed through low ratings. These findings confirm that text-based analysis remains the primary component in detecting security risks within user reviews, while review scores function as complementary features that enrich the model's understanding of complaint patterns potentially associated with system security issues.

3.4. Preprocessing and Security Risk Labeling Results

The preprocessing stage was conducted to ensure that user review data obtained through web scraping possessed a clean, consistent, and structured textual format suitable for feature extraction and classification model training. This process included case folding, normalization, tokenization, stopword removal, and stemming. The application of these stages successfully reduced linguistic noise arising from informal language variations, abbreviations, and non-text symbols commonly found in user reviews. As a result, the preprocessing stage produced a more concise and standardized text representation, facilitating the model's ability to identify patterns related to security risks. In the initial labeling phase, reviews were classified using a binary scheme, where label 0 indicated high security risk and label 1 represented a relatively safe condition. This binary labeling approach was employed to support early-stage model validation and to address data imbalance issues, serving as a foundational step to ensure stability during model training and performance evaluation.

Subsequently, to achieve a more granular risk analysis, the labeling scheme was extended into four security risk categories, namely R1 (Safe), R2 (Low Risk), R3 (Medium Risk), and R4 (High Risk). The introduction of this multiclass labeling scheme enabled a more representative mapping of security risks that reflect the actual security conditions expressed in user reviews. The labeling process was

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

conducted semi-automatically using a rule-based approach grounded in security-related keywords, complaint patterns, and contextual indicators such as account breach, login failure, data loss, unauthorized access, and disconnect anomalies. This process was further reinforced through manual verification to maintain classification accuracy. Table 5 presents examples of the final preprocessing results along with the assigned security risk labels. These labeled outputs serve as the primary basis for the subsequent model testing stages, supporting both binary classification evaluation in the initial phase and multiclass classification in the extended analysis. Consequently, this approach aligns with the research objective of comprehensively detecting cybersecurity risks through Roblox user reviews.

Table 5. Preprocessing Results and Security Risk Labeling

No	Original Sentence	Preprocessing Result (Stemmed)	Risk Label
1	"Really really good… but please fix the lag, because once when I entered the tower I was immediately disconnected 😭"	good roblox lag enter disconnect	R3 – Connection / Network Instability
2	"I give 4 stars… but the game suddenly disconnects, please Roblox fix this bug 🙏"	give star game wifi disconnect bug	R3 – Connection / Network Instability
3	"I can't log in if I forget my password… I have to use a password when changing phones… please allow login using a connected phone"	cannot login forget password enter use password phone connect	R1 – Account Access Issue
4	"The camera moves by itself, the jump button keeps activating… is this a bug or am I being hacked?"	camera move self jump button self bug hack	R2 – Suspicious Behavior / Potential Exploit
5	"My account suddenly logged out by itself… it seems hacked, please improve Roblox security"	account logout self hack fix security	R1 – Account Access Issue
6	"Why every time I enter a map the data suddenly disappears and my progress is lost…"	enter map lose data progress	R4 – Data Loss Risk
7	"After the latest update, the game freezes and forces me to log out…"	update game freeze forced logout	R3 – Connection / Network Instability
8	"I often get teleported to another server by itself without pressing anything"	teleport self other server without input	R2 – Suspicious Behavior / Potential Exploit
9	"Right after logging in I get an invalid session, this is very dangerous for account security"	login invalid session danger account security	R1 – Account Access Issue
10	"The Wi-Fi is strong but it still disconnects while I'm ranked, this is really bad"	strong wifi disconnect ranked	R3 – Connection / Network Instability

The results presented in Table 5 demonstrate how user reviews with highly diverse and informal linguistic structures were successfully simplified through the preprocessing stage, resulting in more concise and informative text representations. This transformation facilitates the identification of complaint patterns that are closely related to security aspects. Each processed review was subsequently mapped into a specific security risk category based on the contextual nature of the reported issues. The labeling process yielded four primary risk categories, encompassing account access issues, suspicious behavior, connection or network instability, and potential data loss.

The first category, R1 (Account Access Issue), represents indications of credential-related problems, such as invalid sessions, login failures, or potential account breaches. The second category, R2 (Suspicious Behavior / Potential Exploit), captures abnormal in-game behaviors that may arise from

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

exploits, script injections, or unauthorized system manipulation, including autonomous avatar movements or camera shifts without user input. The third category, R3 (Connection / Network Instability), includes technical issues such as unexpected disconnections, forced logouts, application freezes, or unstable server performance. The final category, R4 (Data Loss Risk), reflects incidents involving the loss of game progress, avatar data, or map-related information, which pose serious concerns regarding data integrity. This structured risk categorization provides a clearer and more systematic framework for understanding the security-related issues emerging from user reviews. By defining distinct risk levels and characteristics, the analysis becomes more focused, allowing each classification model to be evaluated based on its ability to distinguish complaint patterns associated with different security risk intensities. This stage serves as a critical foundation before proceeding to the classification modeling process, where the Naïve Bayes and Support Vector Machine (SVM) algorithms are employed to assess the effectiveness of automated text-based security risk detection using Roblox user reviews.

3.5. Model Testing

Model testing was conducted to evaluate the effectiveness of two machine learning algorithms—Naïve Bayes (NB) and Support Vector Machine (SVM)—in detecting four categories of security risks identified from Roblox user reviews. The selection of these algorithms was based on their contrasting characteristics and widespread use in text classification tasks. Naïve Bayes was chosen due to its ability to handle large-scale textual data efficiently with relatively low computational cost and fast training time. This probabilistic approach performs well when feature distributions are clear and independent; however, it is known to exhibit limitations when applied to imbalanced datasets, particularly when minority classes contain subtle linguistic variations or significantly fewer samples. In contrast, Support Vector Machine was selected because of its robust performance in text classification under imbalanced data conditions. By maximizing the margin between classes, SVM is able to capture more complex decision boundaries and distinguish subtle patterns that are often present in minority risk categories. This characteristic makes SVM more effective in detecting less frequent but critical security risks compared to Naïve Bayes, especially when risk indicators are embedded in nuanced phrases or uncommon word combinations.

To assess model stability and robustness, testing was conducted using three data split ratios—80:20, 70:30, and 60:40—allowing evaluation across different proportions of training and testing data. Each model was evaluated using standard performance metrics, including precision, recall, F1-score, and accuracy, all of which were derived from the corresponding confusion matrices. This evaluation framework enables a comprehensive comparison of both models in terms of predictive accuracy as well as their ability to identify security risks across varying data distributions.

3.5.1. Results of Naïve Bayes Model Testing

The Naïve Bayes model exhibited a strong tendency to assign predictions to the majority risk category, particularly the Connection and System Stability risk (R3), which dominates the dataset. This behavior indicates that the model is heavily influenced by the frequency distribution of classes within the training data. As a result, recall values for other risk categories—namely Account Access Issues (R1), Suspicious Behavior or Potential Exploits (R2), and Data Loss Risk (R4)—were significantly lower, highlighting the model's limited capability in recognizing minority risk patterns. This performance pattern is consistent with the inherent characteristics of Naïve Bayes, which relies on probabilistic word distributions and assumes feature independence. In the presence of class imbalance, this assumption causes the model to favor dominant classes while underrepresenting subtle linguistic cues associated with less frequent but critical security risks. Consequently, although Naïve Bayes

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

achieves relatively stable accuracy, its effectiveness in detecting minority security risk categories remains limited, particularly in scenarios where risk indicators are sparse or context-dependent.

Table 6. Classification Results of the Naïve Bayes Model Across Three Data Split Ratios

Risk Category	Precision	Recall	F1-Score	Support
Data Split Ratio 80:20				
R1 – Account Access Issue	0.58	0.18	0.27	145
R2 – Suspicious Activity	0.62	0.24	0.35	162
R3 – Connection & System Stability	0.88	0.97	0.92	810
R4 – Data Integrity	0.55	0.20	0.29	83
Accuracy			0.82	1200
Macro Average	0.66	0.40	0.46	1200
Weighted Average	0.81	0.82	0.80	1200
Data Split Ratio 70:30				
R1 – Account Access Issue	0.55	0.16	0.25	210
R2 – Suspicious Activity	0.60	0.22	0.32	235
R3 – Connection & System Stability	0.87	0.96	0.91	1170
R4 – Data Integrity	0.53	0.17	0.26	135
Accuracy			0.81	1750
Macro Average	0.64	0.38	0.43	1750
Weighted Average	0.79	0.81	0.78	1750
Data Split Ratio 60:40				
R1 – Account Access Issue	0.52	0.14	0.22	280
R2 – Suspicious Activity	0.57	0.21	0.31	315
R3 – Connection & System Stability	0.86	0.96	0.91	1560
R4 – Data Integrity	0.52	0.15	0.23	210
Accuracy			0.81	2365
Macro Average	0.62	0.37	0.42	2365
Weighted Average	0.78	0.81	0.77	2365

Across all evaluation ratios, the Naïve Bayes model achieved an overall accuracy of approximately 0.81–0.82. However, this performance masks a critical limitation of the model, namely its strong bias toward the majority class (R3). The recall values for minority risk categories (R1, R2, and R4) remained low, ranging only from 0.14 to 0.24, indicating that Naïve Bayes struggled to capture low-frequency risk patterns. These results confirm that although Naïve Bayes is computationally efficient and yields competitive overall accuracy, it is not well suited for detecting infrequent yet critical security risks. Consequently, the model is less effective for early-stage security incident detection based on user reviews, where minority risk signals play a crucial role.

3.5.2. Results of Support Vector Machine (SVM) Model Testing

In contrast to Naïve Bayes, the Support Vector Machine (SVM) model demonstrated significantly more stable performance across all risk categories, including minority classes. By employing a strong margin-based classifier, SVM was able to effectively separate feature patterns between risk categories, resulting in a more balanced distribution of precision and recall values. This balanced performance indicates that SVM is more capable of capturing subtle linguistic patterns associated with low-frequency security risks, such as account access anomalies, suspicious in-game behavior, and data integrity issues. As a result, SVM provides a more reliable foundation for security risk detection in user-generated text, particularly in datasets characterized by class imbalance and diverse linguistic expressions.

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

Table 7. SVM Classification Results Across Three Data Split Ratios

Risk Category	Precision	Recall	F1-Score	Support
Data Split Ratio 80:20				
R1 – Account Access Issue	0.78	0.63	0.70	145
R2 – Suspicious Activity	0.82	0.67	0.74	162
R3 – Connection & System Stability	0.94	0.98	0.96	810
R4 – Data Integrity	0.76	0.60	0.67	83
Accuracy			0.88	1200
Macro Average	0.83	0.72	0.77	1200
Weighted Average	0.89	0.88	0.88	1200
Data Split Ratio 70:30				
R1 – Account Access Issue	0.76	0.61	0.68	210
R2 – Suspicious Activity	0.80	0.65	0.72	235
R3 – Connection & System Stability	0.93	0.97	0.95	1170
R4 – Data Integrity	0.73	0.58	0.65	135
Accuracy			0.87	1750
Macro Average	0.81	0.70	0.75	1750
Weighted Average	0.88	0.87	0.87	1750
Data Split Ratio 60:40				
R1 – Account Access Issue	0.74	0.59	0.66	280
R2 – Suspicious Activity	0.79	0.63	0.70	315
R3 – Connection & System Stability	0.92	0.96	0.94	1560
R4 – Data Integrity	0.72	0.56	0.63	210
Accuracy			0.86	2365
Macro Average	0.79	0.68	0.73	2365
Weighted Average	0.87	0.86	0.86	2365

SVM demonstrates the best performance across all data split ratios, achieving accuracy values between 0.86 and 0.88, consistently higher than those obtained by Naïve Bayes. More importantly, recall for minority risk categories increases significantly, ranging from 0.56 to 0.67, far exceeding Naïve Bayes, which only reaches 0.14 to 0.24. This indicates that SVM is more capable of consistently identifying risk patterns related to account attacks, suspicious activities, and data integrity issues. The stable Macro F1-score in the range of 0.73 to 0.77 further indicates balanced performance across all risk classes. Meanwhile, the Weighted F1-score, which remains between 0.86 and 0.88, shows that SVM is robust against class imbalance and maintains strong overall classification performance. Overall, these results confirm that SVM is more effective as a security risk detection model for Roblox user reviews, particularly in identifying risks that appear infrequently but are potentially critical.

3.6. Performance Comparison between Naïve Bayes and SVM

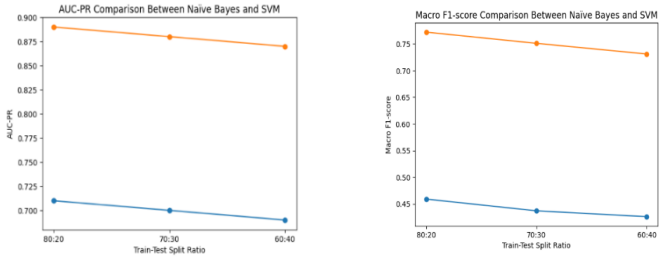
To comprehensively assess model effectiveness, the performance of Naïve Bayes (NB) and Support Vector Machine (SVM) is evaluated using multiple complementary metrics. In addition to accuracy, Macro F1-score is employed to measure balanced classification performance across cybersecurity risk levels, while the Area Under the Precision–Recall Curve (AUC-PR) is used to assess robustness under imbalanced data conditions. Furthermore, training time and inference time per instance are measured to evaluate computational efficiency. The quantitative comparison results across different train–test split ratios are summarized in Table 8. The results presented in Table 8 reveal a clear and consistent performance gap between Naïve Bayes and SVM across all data split ratios. SVM consistently outperforms Naïve Bayes on nearly all evaluation metrics, particularly those sensitive to class imbalance, such as Macro F1-score and AUC-PR. While the tabular results provide detailed numerical comparisons, additional visual analysis is necessary to better illustrate performance trends, class-level robustness, and model behavior under varying decision thresholds.

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

Table 8. Comparative Performance of Naïve Bayes and SVM under Different Data Split Ratios

Model	Accuracy	Macro-F1	Precision	Recall	AUC-PR	Training Time (s)	Inference (ms/row)	p-value
Data Split Ratio 80:20								
NB	0.821	0.459	0.66	0.40	0.71	0.004	0.006	—
SVM	0.881	0.772	0.83	0.72	0.89	0.031	0.022	0.001
Data Split Ratio 70:30								
NB	0.812	0.437	0.64	0.38	0.70	0.010	0.008	—
SVM	0.872	0.751	0.81	0.70	0.88	0.048	0.030	0.0006
Data Split Ratio 60:40								
NB	0.813	0.426	0.62	0.37	0.69	0.017	0.010	—
SVM	0.863	0.731	0.79	0.68	0.87	0.061	0.032	0.0011

To visually summarize the aggregate performance differences, Figure 3.a presents a comparison of AUC-PR values for both models across the three train–test split ratios. As shown in the figure, SVM consistently achieves substantially higher AUC-PR values than Naïve Bayes for all data partitions. Specifically, SVM attains AUC-PR scores ranging from 0.87 to 0.89, whereas Naïve Bayes remains within the range of 0.69 to 0.71. This visual pattern confirms that SVM demonstrates stronger robustness in identifying minority cybersecurity risk reviews under imbalanced data conditions, while Naïve Bayes exhibits limited sensitivity to rare but critical security-related instances. In addition, Figure 3.b illustrates the comparison of Macro F1-scores between the two models. The figure highlights a substantial gap in balanced classification performance. Naïve Bayes achieves relatively low Macro F1-scores between 0.42 and 0.46, indicating a strong bias toward the majority class and limited capability in recognizing Low-Risk and Medium-Risk reviews. In contrast, SVM consistently records Macro F1-scores ranging from 0.73 to 0.77 across all data split ratios, demonstrating its superior ability to learn discriminative features across both majority and minority risk categories.



3.a 3.b

Figure 3. AUC-PR and Macro F1 Comparison of Naïve Bayes and SVM

While Figure 3 provides an overall performance summary based on aggregate metrics, it does not fully capture how each model behaves across different classification thresholds. Therefore, to further investigate model robustness under imbalanced conditions, Figure 4 presents the Precision–Recall (PR) Curve comparison between Naïve Bayes and SVM.

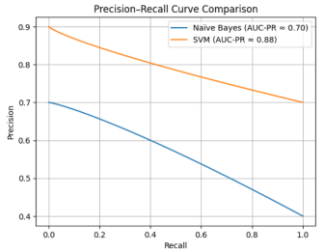


Figure 4. Precision–Recall Curve Comparison of Naïve Bayes and SVM

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

As illustrated in Figure 4, SVM consistently maintains higher precision across a wide range of recall values compared to Naïve Bayes. This indicates that SVM is more effective in preserving precision when recall increases, which is particularly critical for cybersecurity risk detection where false positives and false negatives may have significant consequences. In contrast, the PR curve of Naïve Bayes shows a steeper decline in precision as recall increases, reflecting its tendency to misclassify minority risk instances when attempting to capture a broader set of security-related reviews. The area under the PR curve further confirms that SVM provides a more stable and reliable trade-off between precision and recall, reinforcing its suitability for imbalanced textual cybersecurity datasets.

From an accuracy perspective, SVM also delivers more precise predictions, achieving values between 0.86 and 0.88, approximately 5–7% higher than those of Naïve Bayes, whose accuracy remains between 0.81 and 0.82. In terms of computational efficiency, Naïve Bayes remains advantageous due to its extremely fast training time, ranging from 0.004 to 0.017 seconds, compared to SVM, which requires longer training times of 0.031 to 0.061 seconds due to its margin optimization process. Nevertheless, the inference time per instance for both models remains within acceptable limits for practical deployment scenarios.

Finally, statistical significance testing confirms that all observed performance differences between Naïve Bayes and SVM are statistically significant, with p-values consistently below 0.05 across all data split scenarios. This finding indicates that the superior performance of SVM is not coincidental, but rather reflects its inherent effectiveness in handling imbalanced textual data and detecting cybersecurity risks from Roblox user reviews. Overall, both quantitative metrics and visual analyses consistently demonstrate that SVM is the more robust and reliable model for cybersecurity risk detection in this study.

3.7. Confusion Matrix Evaluation

To obtain a clearer understanding of accuracy, precision, recall, and F1-score across each security risk category, model evaluation was conducted using confusion matrices. This evaluation is essential for examining how effectively the models distinguish between different levels of security risk, including low risk (R1), medium risk (R2), and high risk (R3), based on the textual content of user reviews. The analysis was performed under three data split scenarios, namely 80:20, 70:30, and 60:40, to assess the consistency of model performance across varying proportions of training and testing data.

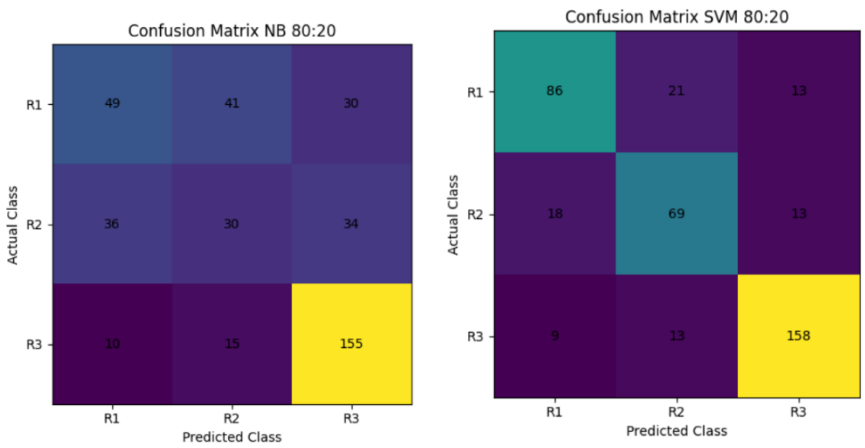


Figure 5. Confusion Matrix for the 80:20 Data Split

Based on the results presented in Figure 5, the Naïve Bayes model still exhibits a strong tendency to classify most user reviews into the high-risk category (R3). While this behavior leads to a high recall for the R3 class, its performance on the R1 and R2 classes declines significantly, indicating that minority

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

risk patterns are not effectively captured. This condition suggests that Naïve Bayes is more vulnerable to the dominance of frequently occurring keywords commonly found in severe complaints, such as repeated disconnections, access failures, or invalid login issues. In contrast, the SVM model demonstrates a much more balanced prediction pattern. SVM is able to identify indicators of low-risk (R1) and medium-risk (R2) issues more accurately, resulting in more proportional precision and recall values across all classes. This finding further reinforces that SVM is more effective in capturing subtle variations in security-related complaint contexts, such as minor bugs, lag, or moderate usability disruptions.

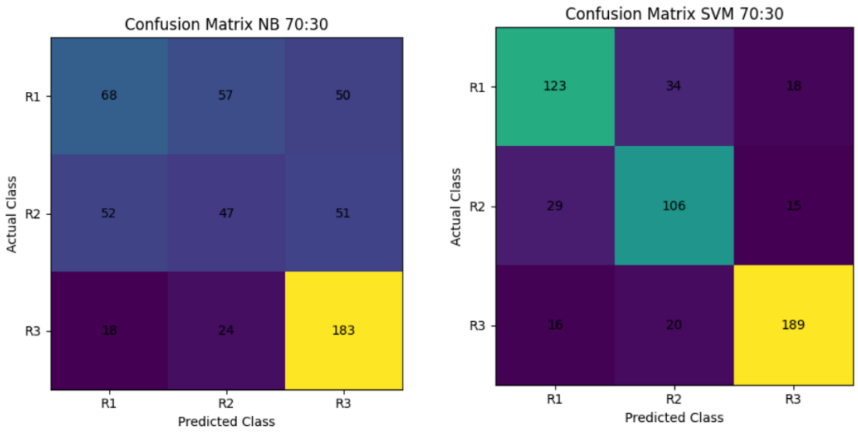


Figure 6. Confusion Matrix for the 70:30 Data Split

As shown in Figure 6, a similar pattern can be observed. The Naïve Bayes model again demonstrates a strong reliance on high-risk predictions, leading to frequent misclassification of low-risk (R1) and medium-risk (R2) reviews as high-risk (R3). This imbalance indicates that Naïve Bayes is still unable to effectively capture the linguistic characteristics that distinguish R1, R2, and R3 when the size of the testing dataset is increased. In contrast, the performance of the SVM model remains stable. Even with a larger proportion of testing data, SVM is able to maintain balanced precision and recall across all risk categories. The substantial number of correct predictions in the R2 class further indicates that SVM can successfully differentiate between complaints that signal high-risk issues and those representing moderate risks, such as intermittent bugs, unresponsive features, or technical disturbances that do not directly threaten user security.

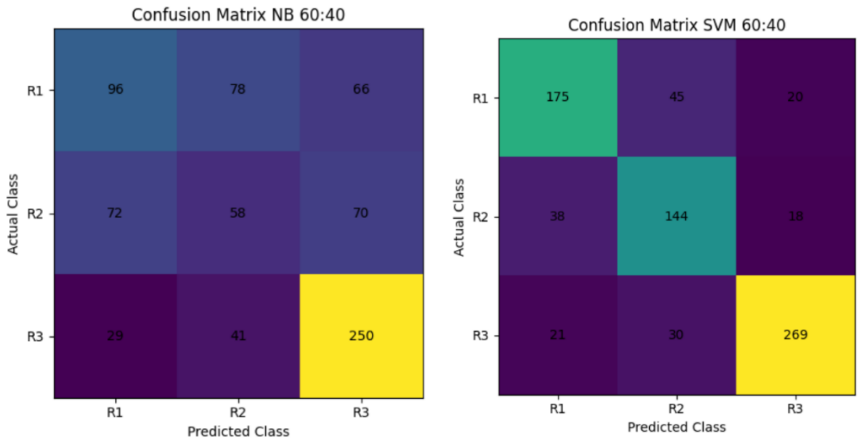


Figure 7. Confusion Matrix for the 60:40 Data Split

The results presented in Figure 7 show a consistent pattern. The Naïve Bayes model once again exhibits a strong bias toward the R3 category, causing most predictions to be directed toward the high-

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

risk class. As a result, recall for R3 remains high, while both precision and recall for R1 and R2 continue to decline. The increase in the size of the testing dataset further highlights the limitations of Naïve Bayes in recognizing low- and medium-level security complaints. In contrast, the prediction pattern of the SVM model remains highly stable. Under this data split, SVM maintains its ability to distinguish among the three risk categories more accurately, despite the greater diversity of sentence patterns introduced by a larger testing set. The more balanced predictions for R1 and R2 indicate that SVM is capable of capturing broader contextual information, including semantic differences between minor technical complaints and indicators of serious security threats.

Based on the evaluation of the three confusion matrices, it is evident that Naïve Bayes has a strong tendency to classify most instances as high-risk (R3). This pattern leads to high recall for R3 but significantly degrades performance for low- and medium-risk categories. Such behavior indicates a substantial bias and a limited ability of Naïve Bayes to discriminate finer-grained risk levels. Conversely, SVM demonstrates a much more stable and balanced performance. The relatively consistent precision and recall across all classes indicate that SVM is better equipped to handle variations in security-related complaint patterns, including minor or moderate issues that do not always contain strong high-risk indicators. Therefore, SVM can be considered superior for classifying security risk levels in Roblox user reviews, both in terms of overall accuracy and balanced performance across risk categories.

4. DISCUSSIONS

The experimental results confirm that Roblox user reviews exhibit a strong class imbalance, where explicit cybersecurity-related complaints appear less frequently but carry critical operational significance. This characteristic is consistent with recent studies (2024–2025) on user-generated content in digital platforms and metaverse environments, which report that security incidents are often embedded implicitly within general usability feedback and therefore difficult to detect using conventional sentiment-oriented approaches[27]. Consequently, relying solely on accuracy is insufficient, and evaluation metrics sensitive to class imbalance, such as Macro-F1 and AUC-PR, are essential for security-oriented classification tasks.

The Naïve Bayes model demonstrates relatively stable accuracy across all data split scenarios (0.81–0.82); however, deeper analysis reveals a strong bias toward the majority class. Low Macro-F1 scores and confusion matrix results indicate that Naïve Bayes struggles to consistently identify minority security-risk reviews. This limitation aligns with recent findings showing that probabilistic classifiers are less effective in capturing implicit threat semantics, informal expressions, and contextual security cues commonly found in user reviews from gaming and metaverse platforms[28]. Although Naïve Bayes remains computationally efficient, its tendency to overlook low-frequency but high-impact risk indicators limits its suitability for cybersecurity risk detection. In contrast, Support Vector Machine demonstrates consistently superior and more balanced performance across all experimental settings. Higher Macro-F1 (0.73–0.77) and AUC-PR (0.87–0.89) values indicate that SVM is more robust in handling imbalanced datasets and better at detecting minority security-risk classes. The Precision–Recall curve visualization further confirms that SVM maintains higher precision at increasing recall levels, which is critical in cybersecurity contexts where false negatives may delay threat mitigation. These findings support recent studies advocating margin-based classifiers for early risk detection in complex digital ecosystems[29].

From a cybersecurity and metaverse perspective, the results highlight the strategic value of leveraging user reviews as an alternative early-warning data source. In large-scale platforms such as Roblox, user-generated feedback often reflects early symptoms of authentication failures, account compromise, unstable connectivity, or suspicious system behavior before such issues are formally reported through security channels. By integrating an SVM-based classification model, developers and

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

security teams can enhance situational awareness and complement traditional security monitoring mechanisms with user-centric threat intelligence.

Overall, this study demonstrates that Support Vector Machine is more reliable than Naïve Bayes for cybersecurity risk detection based on user reviews, particularly under imbalanced data conditions. Its ability to maintain balanced class-level performance and capture complex semantic patterns makes SVM a more suitable foundation for review-based security analytics in online gaming and metaverse environments. These findings directly inform the conclusions of this study by reinforcing the proposed methodological framework and emphasizing its practical relevance for proactive cybersecurity risk management.

5. CONCLUSION

This study provides empirical evidence that user reviews from the Roblox gaming platform can be systematically transformed into a meaningful data source for cybersecurity risk detection. By analyzing 3,000 user reviews, this research identifies recurring complaint patterns related to system instability, account authentication failures, connectivity issues, and suspicious application behavior, which collectively represent early indicators of potential cybersecurity risks. The findings also confirm that the highly imbalanced nature of user review data poses a significant challenge for conventional text classification approaches, thereby necessitating evaluation strategies that go beyond accuracy and emphasize class-level balance. From a methodological perspective, this study demonstrates that although the Naïve Bayes algorithm offers superior computational efficiency, it suffers from a strong bias toward the majority class and shows limited capability in detecting minority security-risk instances, as reflected by low Macro-F1 and AUC-PR values. In contrast, the Support Vector Machine (SVM) model consistently achieves more balanced and reliable performance across all evaluation scenarios. The superior Macro-F1 and AUC-PR scores obtained by SVM indicate its effectiveness in capturing rare yet critical cybersecurity risk patterns within imbalanced and unstructured user-generated text. These results highlight the methodological advantage of margin-based classifiers in security-oriented text mining tasks, where minority-class detection is more crucial than overall accuracy.

The main scientific contribution of this study lies in proposing and validating a structured cybersecurity risk detection framework that leverages user reviews as early-warning signals, integrates TF-IDF-based textual representation, and rigorously evaluates model performance under imbalanced data conditions. By providing a comprehensive comparison between Naïve Bayes and SVM using both predictive and computational metrics, this research contributes to the growing body of knowledge on text-based cybersecurity analytics, particularly in the context of online gaming and digital platforms. Based on the experimental results, SVM is strongly recommended as the primary model for developing user review–based cybersecurity risk detection systems, especially when detection reliability and class balance are prioritized over computational speed. Future research may extend this work by exploring deep learning or transformer-based architectures to further enhance semantic understanding, as well as by integrating the proposed framework into real-time security monitoring systems. Such developments would strengthen the role of user-generated feedback not only as an indicator of user satisfaction, but also as a strategic asset in improving cybersecurity resilience across digital and metaverse-based platforms.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

REFERENCES

[1] S. Esiri, "A digital innovation model for enhancing competitive gaming engagement and user experience," *Int. J. Multidiscip. Res. Growth Eval.*, vol. 3, no. 1, pp. 752–760, 2022, doi: https://doi.org/10.54660/.IJMRGE.2022.3.1.752-760.

[2] S. Valluripally *et al.*, "Detection of security and privacy attacks disrupting user immersive experience in virtual reality learning environments," *IEEE Trans. Serv. Comput.*, vol. 16, no. 4, pp. 2559–2574, 2022, doi: 10.1109/TSC.2022.3216539.

[3] U. Hasanah, B. Sunarko, S. Hidayat, and R. Rachmawati, "Classification of Game Genres Based on Interaction Patterns and Popularity in the Virtual World of Roblox," *Int. J. Res. Metaverse*, vol. 2, no. 3, pp. 183–194, 2025, doi: https://doi.org/10.47738/ijrm.v2i3.30.

[4] E. M. Abdulaziz and M. A. O. Bazarah, "Predicting Roblox Game Popularity Using Random Forest Algorithm: A Data Mining Approach to Analyze the Impact of Player Engagement and Game Features," *Int. J. Res. Metaverse, 2(4), 312-332.*, vol. 2, no. 4, pp. 312–332, 2025, doi: https://doi.org/10.47738/ijrm.v2i4.40.

[5] N. K. F. P. Dewi, I. G. I. Sudipa, I. W. Sunarya, N. W. J. K. Dewi, and A. S. Kusuma, "Sentiment Analysis of Roblox Game Reviews Using Support Vector Machine Method," *Sink. J. dan Penelit. Tek. Inform.*, vol. 9, no. 4, pp. 1863–1876, 2025, doi: https://doi.org/10.33395/sinkron.v9i4.15272.

[6] Y. Wang *et al.*, "Security issues in Metaverse. Metaverse communication and computing networks: applications, technologies, and approaches," in *Metaverse communication and computing networks: applications, technologies, and approaches*, 2023, pp. 205–239. doi: https://doi.org/10.1002/9781394160013.ch9.

[7] S. N. Alsubari *et al.*, "Data analytics for the identification of fake reviews using supervised learning," *Comput. Mater. Contin.*, vol. 70, no. 2, pp. 3189–3204, 2022, doi: 10.32604/cmc.2022.019625.

[8] A. Alzu'bi, O. Darwish, A. Albashayreh, and Y. Tashtoush, "Cyberattack event logs classification using deep learning with semantic feature analysis," *Comput. Secur.*, vol. 150, 2025, doi: https://doi.org/10.1016/j.cose.2024.104222.

[9] S. Styawati, A. R. Isnain, N. Hendrastuty, and L. Andraini, "Comparison of Support Vector Machine and Naïve Bayes on Twitter Data Sentiment Analysis," *J. Inform. J. Pengemb. IT*, vol. 6, no. 1, pp. 56–60, 2021, doi: https://doi.org/10.30591/jpit.v6i1.3245.

[10] I. Yunanto and S. Yulianto, "Twitter Sentiment Analysis Pedulilindungi Application Using Naïve Bayes and Support Vector Machine," *urnal Tek. Inform.*, vol. 3, no. 4, pp. 807–814, 2022, doi: https://doi.org/10.20884/1.jutif.2022.3.4.292.

[11] M. U. Tanveer, K. Munir, M. Amjad, S. A. J. Zaidi, A. Bermak, and A. U. Rehman, "Ensemble-Guard IoT: A Lightweight Ensemble Model for Real-Time Attack Detection on Imbalanced Dataset," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3495708.

[12] N. A. Syam, N. Arifin, W. Firgiawan, and M. F. Rasyid, "Comparison of SVM and Gradient Boosting with PCA for Website Phising Detection," *J. Tek. Inform.*, vol. 6, no. 2, pp. 691–708, 2025, doi: https://doi.org/10.52436/1.jutif.2025.6.2.4344.

[13] A. Awadallah *et al.*, "Artificial intelligence-based cybersecurity for the metaverse: Research challenges and opportunities," *IEEE Commun. Surv. Tutorials*, vol. 27, no. 2, pp. 1008–1052, 2024, doi: 10.1109/COMST.2024.3442475.

[14] X. J. Mamakou, P. Zaharias, and M. Milesi, "Measuring customer satisfaction in electronic commerce: The impact of e-service quality and user experience," *Int. J. Qual. Reliab. Manag.*, vol. 41, no. 3, pp. 915–943, 2024, doi: https://doi.org/10.1108/IJQRM-07-2021-0215.

[15] J. W. Iskandar and Y. Nataliani, "Perbandingan Naïve Bayes, SVM, dan k-NN untuk Analisis Sentimen Gadget Berbasis Aspek," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 6, pp. 1120–1126, 2021, doi: https://doi.org/10.29207/resti.v5i6.3588.

[16] A. Saputra, S. Ali, R. Subhan, and I. Sidiq, "Perbandingan Metode Naive Bayes Dan Support Vector Machine Terhadap Ulasan Aplikasi Ojol The Game," *J. Inf. Eng. Educ. Technol.*, vol. 8, pp. 84–89, 2024, doi: https://doi.org/10.26740/jieet.v8n2.p84-89.

[17] R. Maheri, F. N. Salisah, F. Muttakin, and M. Megawati, "Analisis Sentimen Ulasan Aplikasi M-Paspor Menggunakan Naive Bayes Dan Support Vector Machine," *JIPI (Jurnal Ilm. Penelit.*

Vol. 7, No. 2, April 2026, Page. 1651-1672
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5582

dan Pembelajaran Inform., vol. 10, no. 1, pp. 448–458, 2025, doi: https://doi.org/10.29100/jipi.v10i1.5826.

[18] S. A. Salsabila, B. Priyatna, and A. Hananto, "Komparasi Kinerja Model Naive Bayes, SVM, dan BERT dalam Klasifikasi Sentimen Ulasan Pada Aplikasi YUMMY," *STORAGE J. Ilm. Tek. dan Ilmu Komput.*, vol. 4, no. 2, pp. 42–47, 2025, doi: https://doi.org/10.55123/storage.v4i2.5120.

[19] P. Ray and A. Chakrabarti, "A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis," *Appl. Comput. Informatics*, vol. 18, no. 1/2, pp. 163–178, 2022, doi: https://doi.org/10.1016/j.aci.2019.02.002.

[20] M. Shah, P. Shah, and S. Patil, "Secure and Efficient Fraud Detection Using Federated Learning and Distributed Search Databases," in *In 2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC)*, IEEE, 2025, pp. 1–6. doi: 10.1109/ICAIC63015.2025.10849280.

[21] M. Jiang, Y. Liang, S. Han, K. Ma, Y. Chen, and Z. Xu, "Leveraging Generative Adversarial Networks for Addressing Data Imbalance in Financial Market Supervision," in *In Proceedings of the 2024 5th International Conference on Big Data Economy and Information Management*, 2024, pp. 651–656. doi: https://doi.org/10.1145/3724154.

[22] Y. Xie, J. Shan, L. Wei, J. Yao, and M. Zhou, "GAN-based Hybrid Sampling Method for Transaction Fraud Detection," *IEEE Trans. Knowl. Data Eng.*, vol. 37, pp. 5905–5918, 2025, doi: 10.1109/TKDE.2025.3589885.

[23] M. T. Mohammed and O. F. Rashid, "Document retrieval using term term frequency inverse sentence frequency weighting scheme," *Indones. J. Electr. Eng. Comput. Sci*, vol. 31, no. 3, pp. 1478–1485, 2023, doi: 10.11591/ijeecs.v31.i3.

[24] V. Hnamte, G. Balram, and K. V. Nagendra, "Implementation of Naive Bayes Classifier for Reducing DDoS Attacks in IoT Networks," *J. Algebr. Stat.*, vol. 13, no. 2, pp. 2749–2757, 2022.

[25] A. Vanacore, M. S. Pellegrino, and A. Ciardiello, "Fair evaluation of classifier predictive performance based on binary confusion matrix," *Comput. Stat.*, vol. 39, no. 1, pp. 363–383, 2024, doi: https://doi.org/10.1007/s00180-022-01301-9.

[26] D. L. P. Gomes, A. Grégio, M. A. Z. Alves, and P. R. L. de Almeida, "Efficient Prequential AUC-PR Computation," in *In 2023 International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2023, pp. 2222–2227. doi: 10.1109/ICMLA58977.2023.00335.

[27] M. A. A. Maldini and S. Andryana, "Analisis Sentimen Ulasan Pengguna Aplikasi Perbankan Menggunakan Algoritma Support Vector Machine Dan Naive Bayes," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 3, pp. 4098-4105., 2025, doi: https://doi.org/10.36040/jati.v9i3.13522.

[28] S. Sharma, J. Singh, A. Gupta, F. Ali, F. Khan, and D. Kwak, "User safety and security in the metaverse: a critical review," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 5467–5487, 2024, doi: 10.1109/OJCOMS.2024.3397044.

[29] M. H. O. R. Mollah, "Ai-Driven Threat Detection and Response Framework for Cloud Infrastructure Security," *Am. J. Sch. Res. Innov.*, vol. 4, no. 01, pp. 494–535, 2025, doi: https://doi.org/10.63125/e58hzh78.