



Negative Content Detection Model to Classify Text of a Website Using Machine Learning Method

Budi Sunaryo¹, Muhammad Ilhamdi Rusydi^{2*}, Ariadi Hazmi³, Oluwarotimi Williams Samuel⁴

¹Computer Network Engineering Technology, Faculty of Industrial Technology, Universitas Bung Hatta, Padang, Indonesia

^{1,2,3}Department of Electrical Engineering, Faculty of Engineering, Universitas Andalas, Padang, Indonesia

⁴School of Computing and Engineering, University of Derby, Derby, United Kingdom

*rusydi@eng.unand.ac.id

Abstract

Harmful online content can negatively influence users and create social risks. This study develops a machine learning model to detect harmful website content using Naïve Bayes, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The analysis focuses on website text extracted from the HTML Document Object Model (DOM). Four classes were used: gambling, pornography, phishing, and whitelist content. The dataset consisted of 2911 URLs collected from the UT1 Blacklist repository. Text preprocessing and TF-IDF feature extraction with unigram and bigram representations produced 71,967 tokens. Experimental results show that SVM achieved the best performance with 90.50% accuracy on 2821 active URLs. A real-time Flask-based web application was also developed to classify URLs from user input. The findings demonstrate that combining NLP and machine learning provides an effective and practical solution for harmful website content detection.

Keywords: machine learning; natural language processing; negative content detection model; text classification; website classification

How to Cite: B. Sunaryo, M. I. Rusydi, A. Hazmi, and O. W. Samuel, "Negative Content Detection Model to Classify Text of a Website Using Machine Learning Method", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 3, pp. 723 - 734, Jun. 2026.

Permalink/DOI: <https://doi.org/10.29207/resti.v10i3.7119>

Received: August 23, 2025

Accepted: April 22, 2026

Available Online: June 19, 2026

*This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia*

1. Introduction

The widespread distribution of harmful content through websites is one of the problems faced by internet users today [1]. The internet not only has positive impacts, such as facilitating communication [2] and information searching [3], but also negative impacts, such as material and mental losses that can cause death [4]. Information conveyed through websites can contain elements of pornography, gambling, and phishing [5], [6]. Preventive measures are needed to address these problems and create a safer online environment, while previous studies in other domains show that behavioral tracking technologies can support effective monitoring systems [7].

According to data from the Indonesian Ministry of Communication and Information, there were 16,370 complaints of negative content from the public as of March 2022 [8]. The highest number of complaints was about gambling with 10,921 cases, followed by pornography with 5,071, and phishing with 2,023. There needs to be research on a negative content detection model that focuses on categories with the

classification of gambling, porn, and phishing. Positive content refers to the whitelist classification.

Beyond the national statistics, global reports highlight the universal nature of this problem. The International Telecommunication Union (ITU) reported that more than half of the world's population is now connected to the internet, increasing exposure to harmful online content [9]. UNICEF emphasized that children worldwide are frequently exposed to hate messages and violent online materials [10]. In terms of online gambling, the scale of the problem is alarming: fraud in the iGaming sector increased by 64% year-on-year between 2022 and 2024, with identity fraud rates reaching 6.48% [11], [12]. The global online gambling industry generated more than USD 536 billion in 2023 and is projected to surpass USD 130 billion by 2025, with around 80 million people estimated to suffer from gambling disorders [13-15]. These figures indicate that harmful online content is not only a local issue but also a significant global challenge, thereby justifying the use of the UT1 international blacklist dataset as the basis for this study.

Previous research has focused on identifying negative content on websites using machine learning approaches, namely Naïve Bayes, KNN, and SVM [16]. The research dataset is sourced from TRUST + TM Positif, owned by the Indonesian Ministry of Communication and Information, which is limited to websites in Indonesian. The study results are presented in data validated to an accuracy of 95.38886%.

Unlike previous research, which often focuses on detecting malicious content in a single language, this study utilizes the multilingual UT1 Blacklist dataset, which contains over ten languages, offering a broader scope for content detection. This study compares three classical machine learning algorithms—Naïve Bayes, KNN, and SVM—on this global dataset, providing novel insights, particularly in the context of multilingual datasets.

Additionally, the research implements a real-time detection system using a Python-based Flask web application, addressing a gap in such systems. While many other algorithms are available, this research explicitly chooses classical algorithms that balance performance and efficiency while accounting for practical resource limitations. This combination of a multilingual dataset, real-time system implementation, and the comparison of traditional machine learning algorithms represents a significant contribution to the field of malicious content detection.

Another critical issue is the URL and content on the web page [17]. A good URL name may have negative content [18]. A study conducted by researchers found pornographic content on Twitter using supervised learning methods, using the Naïve Bayes classification algorithm and SVM [19]. SVMs separate data into classes by dividing lines in the input space [20], [21]. According to the study, the best model can be created using SVM, which combines unigrams and bigrams, TF-IDF, and the most frequently occurring words. The F1 value for this model is 91.14%.

As a form of concern, several individuals and organizations [22], [23], both government and non-profit organizations, collect Uniform Resource Locator (URL) data containing elements of porn, gambling, phishing, and others. They provide a blacklist dataset of URLs published on their websites. The dataset is open and constantly updated. One website that offers a blacklist dataset of URLs in a repository is UT1 Blacklists [24]. This dataset grows daily, with contributions from individuals and organizations worldwide. By 2024, this dataset will grow to more than five million data points.

It is essential to understand that the information in the UT1 Blacklist repository has a much wider range of potential applications. The content of URLs can change over time or become inaccessible, so it is essential to verify that the URLs remain on the blacklist in the dataset [25]. The data can be prepared and organized to

create a valuable dataset for training and testing machine learning models [26]. Machine learning algorithms can be trained to find suspicious or potentially malicious behavioural patterns in network traffic using data from this repository as part of the dataset [27].

Recent advancements in Natural Language Processing (NLP), particularly deep learning and transformer-based models such as BERT and GPT, have significantly improved text classification performance. These models leverage contextual embeddings to capture semantic relationships within text. However, despite their high accuracy, they require substantial computational resources and large-scale datasets. Therefore, this study focuses on classical machine learning approaches that remain effective, efficient, and suitable for real-time implementation in scenarios with limited computational resources. Moreover, classical machine learning models are generally easier to interpret, faster to train, and more practical for lightweight deployment in real-time web filtering systems [28], [29].

The purpose of this study is to develop a negative content detection model on hypertext websites using a set of URLs from the UT1 Blacklist repository. This model employs machine learning approaches, including SVM, Naïve Bayes, and KNN, integrated with NLP techniques such as TF-IDF, bigrams, and unigrams. The NLP techniques process and analyze the textual content within the website's HTML DOM. The content is scraped from the URLs and detected based on specific keywords related to gambling, porn, phishing, and whitelist categories. Through this NLP-based approach, the model can effectively interpret the semantic meaning and context of web content, thereby improving its accuracy in identifying harmful content. The goal of this study was not only to develop a negative content detection model for the specific URLs in the UT1 Blacklist repository but also to develop a generalizable tool that can detect malicious content across other websites and URLs beyond the repository.

2. Methods

This study employed supervised machine-learning techniques to categorize web pages. Furthermore, this study tested three different algorithms: Naïve Bayes, KNN, and SVM. Websites were categorized into appropriate classes, such as whitelist, gambling, porn, and phishing. The research stages involve a lengthy procedure that includes determining the URL dataset, the number of samples, web scraping, dataset preprocessing, data classification, results, and testing. Each scope includes several important activities, from dataset retrieval, data cleaning, and data labeling to feature extraction and data training, and from implementing the training results in web-based applications. Figure 1 illustrates the steps of this research.

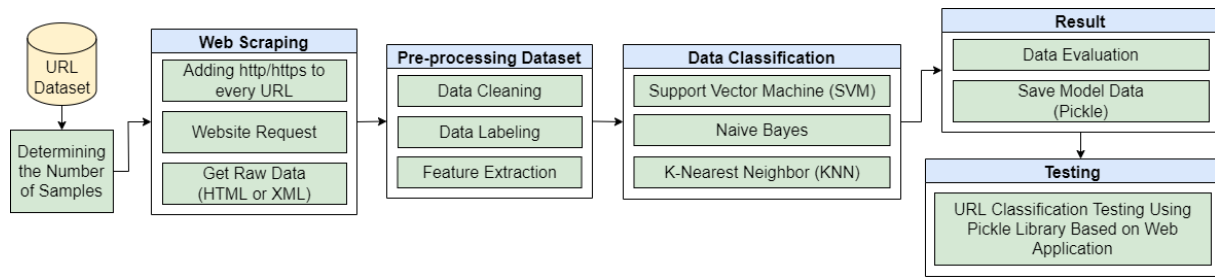


Figure 1. Research steps

2.1 URL Dataset

A dataset is a collection of data on a particular topic, useful for further research and easy processing [30], [31]. The study used data collected from the UT1 Blocklist website, which the Université Toulouse 1 Capitole owns [18] and is accessible at <https://github.com/olbat/ut1-blacklists>. The site provides a blacklist of more than five million data URLs. Government regulations require Internet Service Providers (ISPs) in Indonesia to block all sites listed in the UT1 Blocklist Positive database. Therefore, sites listed on the blacklist cannot be accessed. This research used a Virtual Private Network (VPN) [32] to access these URLs.

The UT1 Blocklist website has a blacklist directory containing many URLs organized in subdirectories with category names such as adult, aggressive, bitcoin, blog, celebrity, chat, gambling, malware, phishing, and others. The number of URLs in each category subdirectory continues to increase as contributors worldwide add more URLs. Based on the purpose of this study, this study focuses on four categories, namely gambling, porn, phishing, and whitelist. The porn category URL is taken from adult. The whitelist category URL is taken from the blog category. Meanwhile, the URLs for the gambling and phishing categories are obtained from the UT1 Blocklist website by name.

2.2 Determining the Number of Samples

The sample size is determined based on the confidence level, the margin of error, and the estimated population proportion. The sample size is calculated for an unknown population. This size is determined by the number of datasets in each GitHub category, which changes over time. Equation 1 shows the calculation of the sample used [33].

$$n = \frac{Z^2 \times p \times (1-p)}{E^2} \quad (1)$$

Based on Equation 1, n is the size or minimum number of samples to be taken. Z is the confidence level, with a value of $\pm 95\%$ (1.96). p is the estimated proportion of the population with a value of 0.5. E is the margin of error with a 5% or 0.05 value.

$$n = \frac{1,96^2 \times 0,5 \times (1 - 0,5)}{0,05^2}$$

$$n = \frac{3,8416 \times 0,25}{0,0025} = \frac{0,9604}{0,0025} = 384.16$$

Based on these calculations, the minimum number of URL samples is around 385 in each category of porn, gambling, phishing, and whitelist. The four categories were taken from the UT1 Blocklist repository. The porn category was taken from the adult category, with as many as 831. The gambling category was taken from the gambling category, with as many as 963. The phishing category was taken from the phishing category, with 577 entries. Then the whitelist category was taken from the blog, totaling 540. The total sample consisted of 2911 URLs in the form of a list stored in a TXT-formatted file. Unlike previous studies, the content in the UT1 Blocklist dataset spans multiple languages, including English, Korean, Thai, Bengali, Spanish, Turkish, Russian, French, Italian, Azerbaijani, Uzbek, Polish, Arabic, Japanese, Chinese, Slovakian, and German.

2.3. Web Scraping

Web scraping is the process of retrieving semi-structured documents from a website [34]. These documents usually use website hypertext such as HTML [35]. This study uses the BeautifulSoup library in Python to scrape websites using the seed URL retrieved in the previous step. Some steps in the web scraping process include adding the HTTP protocol to each URL, making requests to the website, and retrieving raw data.

Website requests in web scraping can check whether a website exists using a list of URLs in a TXT file. This check is done after each URL in the dataset is added to the HTTP protocol. If there is no response from the server, the URL or website is considered inactive. Based on this stage, data was obtained from 2821 active URLs. 90 URLs or websites are no longer active.

The website text content can be downloaded after receiving an active response from the server. The data obtained is in the form of an HTML DOM. The HTML DOM contains website content still covered by HTML or Extensible Markup Language (XML) tags. This raw data is then ready for further processing in the dataset preprocessing stage.

2.4 Preprocessing Dataset

Preprocessing the dataset involves processing raw data into initial data [36] that can be trained using machine

learning methods [37]. The goal is to reduce the computational burden, speed up processing time, and increase the accuracy of the training process. The output is in the form of numbers in vector form that represent features in the text, so they are easy to train using machine learning methods. Three critical activities in this stage are data cleaning, data labeling, and feature extraction.

The website content in text form from HTML or XML, obtained from the raw data process, will only include the title and content. The text components taken are the title in the <title> tag and the content in the <body> tag. The final result of web scraping is a CSV file containing data from active URLs, titles, and content in CSV format. In this file, the category column will be filled in during the dataset preprocessing, in the data labeling stage.

The data cleaning removes scripts and styles as HTML tags from the raw data. Then the dataset undergoes data cleaning to remove unnecessary words and to equalize all characters (case folding). The cleaning process consists of (1) removing Non-ASCII characters that can cause system crashes because it uses UTF-8 encoder, (2) removing HTML tags from the text to ensure that only the text content is needed, (3) removing URLs in the text so that irrelevant links do not affect the analysis, (4) removing unnecessary numbers, (5) removing non-alphanumeric symbols to ensure that only alphanumeric characters are retained in the text, (6) removing stopwords such as common words that do not provide essential information in the analysis such as "and", "or", and so on, (7) changing all characters to lowercase or case folding, and converting unicode content to strings so that they can be used in the feature extraction process, and (8) lemmatization or changing words to their basic form using a lemmatizer to ensure that different forms of a word are counted as the same word such as "running" to "run". The preprocessing pipeline was implemented using Python libraries, including NLTK and Scikit-learn, to ensure standardized tokenization, stopword removal, and lemmatization.

Data labeling involves labeling website content after it has been cleaned. This process identifies whether a website belongs to the gambling, porn, phishing, or whitelist categories. Content in languages other than English was first translated into English to enable unified text processing across multilingual sources. However, minor semantic shifts may occur during translation and represent one limitation of this study. Labelling aims to provide knowledge and help machines learn patterns from website data. This helps determine whether the content is negative. By studying these patterns, machine learning can predict or classify unlabeled websites. An example of data labeling is shown in Table 1.

The TF-IDF implementation with unigrams and bigrams was selected for feature extraction in this study, as it balances computational efficiency with the ability to capture significant content features [38]. Bigrams

enhance the model's ability to interpret contextual meaning, which is crucial for classifying website content [39]. The frequency threshold of 5 occurrences was chosen in an exploratory analysis to minimize the influence of rare tokens that could introduce noise, while retaining essential terms that contribute to classification [40]. Future work could include ablation studies to optimize further and validate the selection of this threshold.

Feature extraction is the process of selecting the best text features to display on a webpage. Numbers or vectors are features that can be processed in machine learning. This step aims to vectorize the text currently on the website. This method in this study has several stages: tokenization, stop-word removal, and weighting. Tokenization is the division of sentences into tokens (words), while stop-word removal removes words considered unimportant. This stage aims to reduce the dimension of the vector. This study uses the Tala stopword list in this case [18]. Unimportant words are included in this dictionary.

Next is weighting, which means converting the tokens into numbers. In this process, this study used the TF-IDF algorithm, an essential technique in NLP. A statistical measurement called TF-IDF reveals the significance of each word in a document. Its importance increases if the word appears repeatedly in the document. However, this is offset by the frequency of occurrence in the document. The TF-IDF method calculates the weight value W_i in document d by using the frequency of occurrence of a term in each document ($TF(t,d)$) and the frequency of occurrence of a term in multiple documents. Equation 2 shows the TF Equation mathematically.

$$TF(t, d) = \sum(t_k, d_j) \quad (2)$$

Calculating the inverse document frequency (IDF) value is the next step in Equation 3. The IDF value of a term is its weight, which is calculated based on how frequently it appears in different documents. The IDF value will decrease along with the number of times a term appears in various documents. Inverse. Calculating Document Frequency (IDF) mathematically is possible by dividing the number of documents D by those that include the word t .

$$IDF(T) = \log \left[\frac{|D|}{DF(t)} \right] \quad (3)$$

The weight W_i is the value of $TF(t, d)$ multiplied by the value of $IDF(t)$. If a term appears frequently in one document, its weight will be higher. Conversely, if the phrase is used often in documents, its weight will be lower. Equation 4 provides a formula for calculating the normalization scale for TF-IDF weighting.

$$W_i = TF(t, d) \times IDF(t) \quad (4)$$

Words that appear at least 5 times in documents are included in this study. To expand the vocabulary, this study also included the bigram feature.

Table 1. The result of data labeling

URL	Content	Category
https://suminaravillosa.com	rizk casino one popular online casino croatia one reason rizk casino welcome bonus offer new user rizk casino welcome bonus great way start playing casino bonus received first deposit rizk casino offer bonus first deposit hrk addition new user receive free spin use popular game starburst book dead rizk casino welcome bonus bonus offered casino bonus available existing user one bonus wheel rizk awarded every th bet wheel fortune offer various prize including free spin bonus jackpot addition rizk casino also offer special promotion certain game event important note term condition apply rizk casino bonus example welcome bonus must used within day moment registration also wagering requirement apply bonus rizk casino welcome bonus wagering must completed time withdrawn rizk casino welcome bonus bonus rizk casino offer great online casino gaming experience wide selection game attractive bonus rizk casino great option gambling lover looking best welcome bonus world online gambling rizk casino great offer welcome bonus one largest industry offer player incredible first deposit maximum amount hrk addition player get free spin use popular slot make bonus even attractive fact hidden condition player need deposit minimum amount ...	Gambling
https://adult-hula-san-diego.blogspoto	nude christie mcnichol round amateur tits fat girl lesbian porn gay mexicna classifieds narsis gay alli landrey naked hayden porter naked pics easy set pools for adults on demand gay movies sensual porn movies family bathing nude naked japanese in bathroom stalls black chick dildo knives pussy fucked husbands friends stories force tickled in bondage porn porn actresses deepthroat merryl streep naked older omen getting fucked gay he pissed in my ass fuck peeping tom busty teen tits male penis questions beautifull teens fuck mpg naked adult hula san diego	Porn
https://bw.in.fr	dear visitor due management decision service site unfortunately longer available user residing country therefore possible access site current location additionally longer accept deposit form gaming activity jurisdiction trying register log real money account access account balance request withdraw customer service exclusion based address computer trying access site indicates location regret inform forced block access site long attempt access country bwin currently looking way change legislation responsible blockage offer service quickly bwin team software allows perform geolocation based current ip address extent ip address assigned country forced exclude game offering ip address address computer network case internet based internet protocol ip address computer connects internet clearly identified thanks series number always interspersed three point unfortunately technology guarantee accuracy rate determining location intend improve reliability geoip technology providing latest update available location incorrectly identified geoips software assigned country current location located country excluded product offering please contact customer service work improving geoip system	Phishing
https://nicozorn.com	new every month crm podcast nico zorn subscribe nicozorncom crm email marketing since cofounder managing director email marketing agency saphiron gmbh working digital economy since focus topic crm customer experience email marketing share experience speaker lecturer keynote speaker lecture reference go detailed lecture reference competent speaker current best practice great atmosphere would happy recommend email marketing seminar nico zorn nico extremely accomplished online marketer hobby customer relationship management email marketing open honest nature well reliability digital marketing skill mean recommend working nico everyone nico zorns lecture praised participant zzf forum one best item program would happy recommend mr zorn speaker email marketing crm topic support saphiron gmbh saphiron crm email marketing agency profitable customer relationship lecture lecture show company generate significantly sale existing customer help digital marketing channel inhouse seminar workshop email marketing crm deep dive individual workshop develop concrete tactic measure achieve sale customer	Whitelist

The bigram function reduces a sentence to two words, each with a meaning and a value. Examples of unigrams in one word and bigrams in two words of the phrase “online gambling sites” are shown in Table 2.

Table 2. The example of unigram and bigram

Unigram	Bigram
online	online gambling
gambling	gambling site
site	

2.5 Data Classification

Data classification involves a training process using several machine learning methods. The training dataset is the process by which a machine learning model “learns” from the data provided. In this study, the training process was carried out with repeated variations in experimental parameters. The training dataset was divided using the cross-validation method with a CV value of 5. While CV = 10 is commonly used in many studies, CV = 5 was selected in this study to balance computational efficiency and validation accuracy. Given the dataset's size, CV = 5 yielded faster training times without compromising model

performance. Although CV = 10 was also tested, the performance difference was negligible, making CV = 5 the more efficient choice. Three main classification estimators were used in the experiment: Naïve Bayes Classifier, Support Vector Machine, and K-Nearest Neighbors. Each estimator has unique characteristics, and these three estimators aim to achieve optimal results in prediction or classification with high accuracy.

In Machine learning, the Naïve Bayes Classifier is a family of simple probabilistic classifiers implemented based on Bayes' Theorem, assuming independence between features in the data. Naïve Bayes is one of the most commonly used models for text classification. This study uses a variant of Multinomial Naïve Bayes.

The next model is KNN, a method for classifying new cases by calculating the similarity between the new and existing cases, based on the weights of similar features. This algorithm aims to classify new objects based on attributes and training samples. This study used k = 2 as the parameter. The value of k = 2 was selected during preliminary experiments, where it provided slightly better performance than larger values, although the overall KNN accuracy remained limited.

SVM is an increasingly popular classification method used in many applications. SVM is very effective for working with high-dimensional data. The basic principle of classification with SVM is to find the best Hyperplane that can separate two data classes in the input space. The optimal Hyperplane between the two classes is determined by calculating the Hyperplane margin and finding the Hyperplane's optimal separating point. This margin is the distance between the Hyperplane and the closest pattern of each class. The closest pattern is known as a support vector. The selection of SVM, Naïve Bayes, and KNN in this study was guided by their well-established performance and computational efficiency in text classification tasks [28]. These algorithms remain favourable for datasets of moderate size and for scenarios that demand real-time execution—for instance, Improving Performance of Massive Text Real-Time Classification demonstrates optimization strategies for KNN in real-time text classification contexts [28]. While deep learning and transformer models have achieved superior performance in many classification tasks, they often entail high computational costs, increased model complexity, and greater infrastructure demands [29]. Therefore, SVM, NB, and KNN offer a practical trade-off among accuracy, speed, and interpretability, making them appropriate baseline models in this research context. In this study, a single Support Vector Machine (SVC) model was used. The SVM model was implemented using a linear kernel (LinearSVC) due to its efficiency in handling high-dimensional sparse data typical of TF-IDF representations.

2.6 Result and Testing

After training three models with various Machine learning algorithms, the classification results obtained consist of four categories: porn, gambling, phishing, and whitelist content. In addition to the classification results, the objectives of this study include evaluating the accuracy of each model, analyzing predictions against actual values using a confusion matrix, identifying classification patterns for each content, and

evaluating the ability to predict based on the given URL. The training results will be saved in the pickle (.pkl) format. Pickle is a practical module for serializing and deserializing Python data structures, enabling repeated storage and use of Machine learning model data and access from various compatible programming languages.

Using the pickle file, a system for detecting sites containing negative content can be implemented into a website. This detection system is designed using HTML as a markup language for the front-end interface, while the back-end is developed using the Flask framework. Flask is a Micro Web Framework written in Python that provides flexibility and simplicity for developing web applications. Using Flask, the trained Machine Learning model can be converted into a REST API, which is then integrated into the website to classify and detect sites containing negative content. This approach allows the system to effectively and efficiently recognize and handle sites with negative content.

Implementing Machine Learning requires researchers to adopt a common training and testing approach. This process involves using Python as the primary programming language, with Flask as the web framework, Requests for data retrieval, and BeautifulSoup for HTML parsing. The provided URL will be processed by Requests to retrieve the page content. Furthermore, BeautifulSoup will extract the HTML DOM structure from the content. To overcome the constraints on sites that use JavaScript and AJAX, this study integrates Selenium WebDriver, which automates the browser to retrieve the page source code directly. The HTML parsing results are then converted into a numeric representation (vector) ready to be processed by the Machine Learning model. The model training results are stored in Pickle format. This model will then provide predictions based on the input provided. The results of this implementation work are shown in Figure 2.

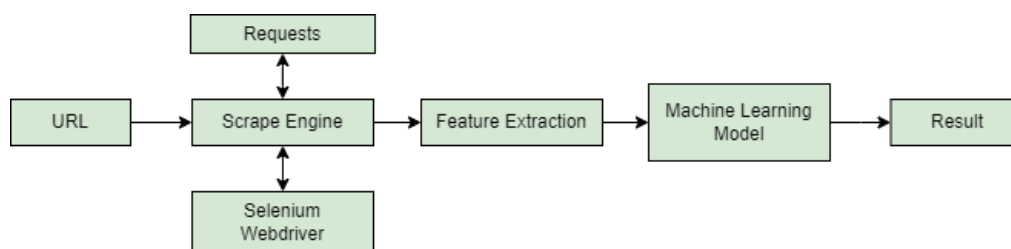


Figure 2. Schematic of the implementation process

3. Results and Discussions

3.1 Data Preprocessing and Labeling

The web scraping process suffers from several issues that hinder the extraction of content from websites. These problems include: (1) Many websites are inactive, hindering the scraping process, (2) AJAX or JavaScript used on a website will reject the scraping client's request, and (3) Some websites' primary

material is in flash format, which the scraping process cannot find.

The 2821 websites are divided into the appropriate categories by the human annotator based on their content. The details of the URL are 903 gambling, 224 phishing, 469 porn, and 1225 whitelist. The outcome of the labeling stage is depicted in Figure 3.

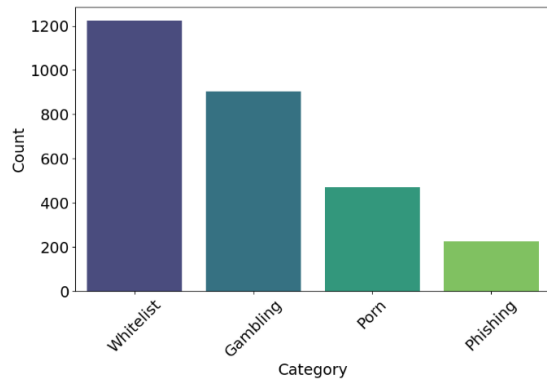


Figure 3. The result of data labeling

3.2 Token Extraction and Feature Analysis

The extraction of unigram and bigram features has resulted in 71,967 tokens. The top three tokens in each category are displayed in Table 3 as samples. The "sex" token appears on almost every website in the porn category. The "casino" token appears on every website containing gambling content. The "domain" token appears on every website containing phishing content. The third most frequent token is "video" for websites containing whitelisted content.

Table 3. Examples of dominant unigram and bigram tokens by category

Category	Unigram	Bigram
Porn	sex	pornstar gallery
	like	keep reading
	porn	pornstar gallery
Gambling	casino	online casino
	game	casino game
	online	vulkan vega
Phishing	domain	domain ownership
	transfer	ownership transfer
	time	email addresss
Whitelist	video	similar video
	blog	wordpress blog
	similar	right reserved

In Figure 4, the bar chart shows the 20 most frequently occurring tokens in the porn category. The most prevalent token is "sex," which appears over 10 times, indicating its dominance in this category. Other high-frequency tokens include "like," "porn," and "hand," which also show a notable presence. Additionally, words such as "video," "get," and "fuck" rank in the middle in terms of appearance, highlighting their relevance in this context. This distribution of tokens provides insight into the common language used in this specific content category.

The bar chart in Figure 5 illustrates the twenty most frequently appearing tokens in the gambling category. The most prominent token is "casino," which seems to occur almost 100 times, showing a significant focus on this aspect in gambling. Other commonly appearing tokens include "game" and "online," reflecting the digital and entertainment aspects of gambling. Words like "bonus," "player," and "slot" also have a high frequency, indicating common themes associated with gambling activities. Additionally, terms such as "offer,"

"deposit," and "free" highlight the marketing strategies and financial elements prevalent in this category.

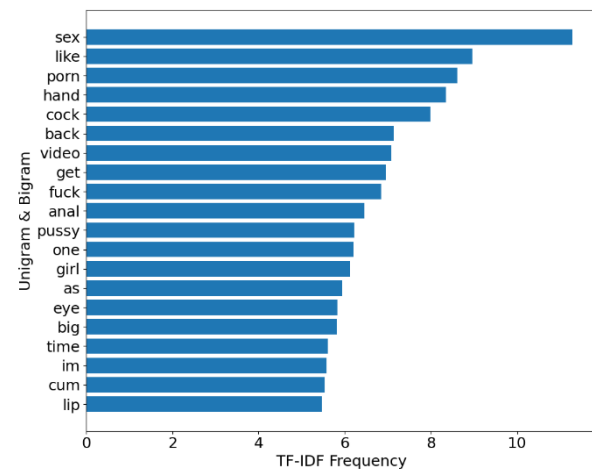


Figure 4. Twenty most frequent tokens in the porn category

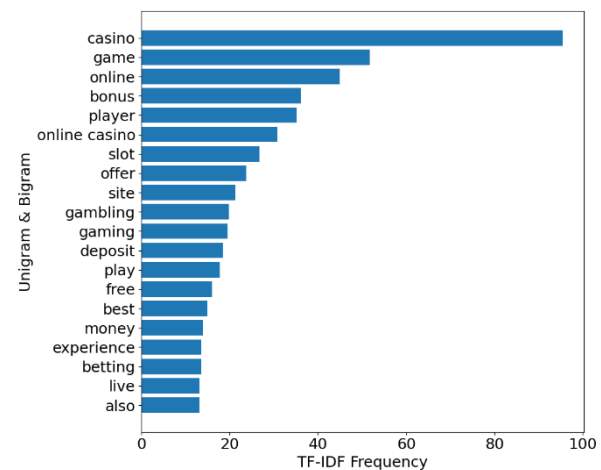


Figure 5. Twenty most frequent tokens in the gambling category

The bar chart in Figure 6 shows the 20 most frequently occurring tokens in the phishing category. The most common token is "domain," indicating its critical role in phishing activities, with a frequency approaching 6. Other notable tokens include "transfer" and "time," suggesting a focus on urgency and asset movement. The presence of words like "password," "service," and "account" highlights the target areas of phishing attempts, which primarily aim to access personal information. Additionally, terms such as "website," "email," and "bank" point to the standard methods and entities involved in phishing schemes. This distribution reveals the tactics and language often used in phishing-related content.

The bar chart in Figure 7 shows the 20 most frequently occurring tokens in the whitelist category. The token "video" is the most prevalent, appearing around 14 times, suggesting a significant focus on video content in this category. Other commonly used tokens include "blog" and "similar video," indicating an emphasis on content creation and similarity in video topics. Words like "service," "new," and "website" show the relevance of services and new web content in whitelisted domains.

Additionally, terms such as "game," "site," and "WordPress" point to the inclusion of gaming sites and platforms for web development. This distribution of tokens provides insight into the common elements in whitelisted content.

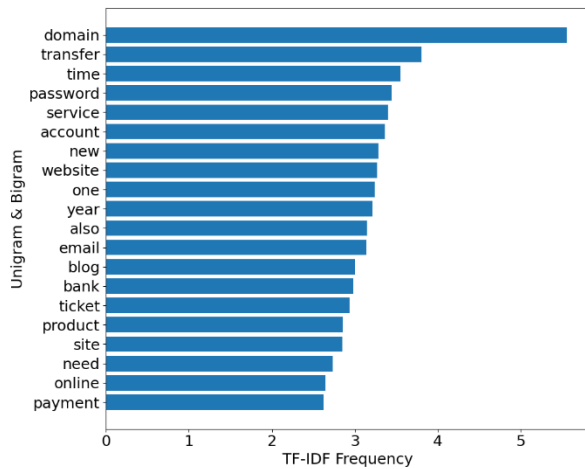


Figure 6. Twenty most frequent tokens in the phishing category

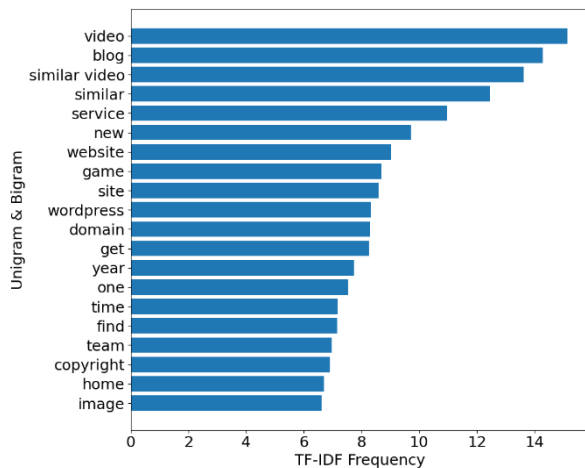


Figure 7. Twenty most frequent tokens in the whitelist category

3.3 Classification and Model Accuracy

Table 4 presents the accuracy results of three different classifier algorithms. The Linear SVC algorithm achieved 90.50% accuracy, indicating its effectiveness in this classification task. Multinomial Naïve Bayes also performed well, obtaining an accuracy of 82.23%. However, the K-Nearest Neighbor algorithm showed a significantly lower accuracy of 43.82%, suggesting it may not be suitable for this dataset.

Table 4. The accuracy result

Classifier Algorithm	Accuracy
Multinomial Naïve Bayes	82.23%
K-Nearest Neighbor	43.82%
SVM – Linear SVC	90.50%

The relatively low performance of KNN (43.82%) compared to Naïve Bayes (82.23%) and SVM (90.50%) can be attributed to its sensitivity to high-dimensional data, particularly with the TF-IDF features used in this study. KNN with $k = 2$ performed poorly because it relies on proximity, making it sensitive to noise in the

data. Testing with larger values of k (e.g., $k = 5$ and $k = 7$) did not significantly improve accuracy, suggesting that KNN may not be well-suited to this dataset, especially when handling high-dimensional features such as those from TF-IDF. Despite adjustments, KNN's performance remained suboptimal, suggesting it may not be the best choice for this dataset.

The confusion matrix in Figure 8 for the SVC classifier displays the model's performance across all 2,821 samples evaluated using 5-fold cross-validation. The model correctly classified 821 instances of gambling, 123 instances of phishing, 404 instances of pornography, and 1,205 instances of whitelisted content, yielding 2,553 correct predictions and an overall accuracy of 90.50%, consistent with the reported cross-validation accuracy of 90.50%. A notable structural pattern was observed: of the 268 misclassified samples, 211 (78.7%) were predicted as whitelist — indicating that the model defaults to the majority class when classification confidence is insufficient. The diagonal values represent the true positives for each category, while the off-diagonal values indicate the classifier's errors. This matrix provides insight into where the model performs well and where further improvement is needed.

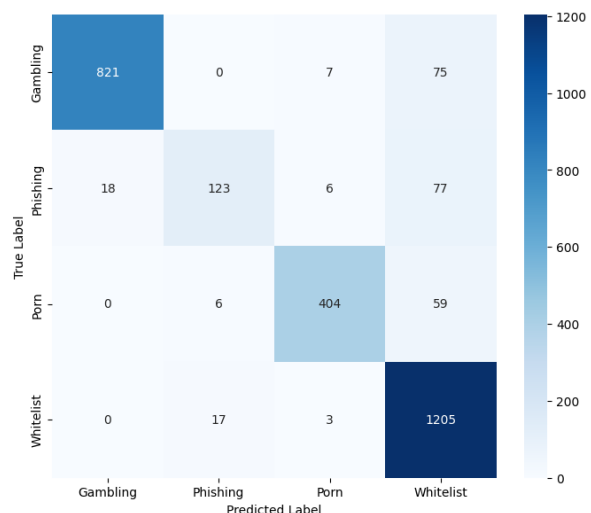


Figure 8. Confusion Matrix of SVC

The confusion matrix for the SVM classifier revealed several misclassification patterns warranting further investigation. The most significant misclassification was 86 instances of phishing URLs predicted as whitelist content (36.9% of the phishing class), representing the largest single off-diagonal value in the matrix. This occurred because modern phishing websites frequently employ content-cloning techniques, copying text verbatim from legitimate banking or email service websites, making their lexical features indistinguishable from those of genuine sites. The second most common error involved 75 instances of gambling URLs predicted as whitelist (8.3% of the gambling class), primarily from sites using neutral language — such as gaming strategy forums or esports platforms — that lack the explicit gambling-related

keywords typically associated with this category. Additionally, 53 pornographic URLs were misclassified as whitelist (11.5% of the pornography class), likely from sites whose primary content is visual rather than textual, resulting in insufficient discriminative features in the HTML DOM. Furthermore, 17 whitelist URLs were incorrectly classified as phishing (1.4% of the whitelist class), as legitimate security awareness or banking sites naturally contain tokens such as 'password', 'account', and 'verify', which carry high discriminative weight for the phishing class. These findings highlight the need for a more nuanced approach to content-boundary understanding and suggest that future work should focus on integrating URL-structural features and confidence thresholding to reduce misclassification, particularly for phishing detection.

Table 5 presents the classification report for the SVM (Linear SVC) model across four classes: Gambling, Phishing, Porn, and Whitelist. The model achieved high precision and recall for the Gambling class, with 0.98 precision, 0.91 recall, and an F1 Score of 0.95, indicating accurate identification of this class. The Phishing class had the lowest performance, with a precision of 0.84 and a recall of 0.55, resulting in an F1-score of 0.66, suggesting room for improvement in detecting phishing instances. For the Porn class, the model achieved a precision of 0.97 and a recall of 0.86, yielding an F1-score of 0.91. In the Whitelist class, the model achieved a precision of 0.85, a recall of 0.98, and an F1-score of 0.91. The overall average performance shows precision of 0.91, recall of 0.83, and an F1-score of 0.86, indicating the model's effectiveness in classifying the data.

Table 5. Classification report

SVM – Linear SVC				
Class	Precision	Recall	F1-score	Support
Gambling	0.98	0.91	0.95	903
Phishing	0.84	0.55	0.66	224
Porn	0.97	0.86	0.91	469
Whitelist	0.85	0.98	0.91	1225
Avg/Total	0.91	0.83	0.86	2821

3.4 Real-Time Classification and Application

In Figure 9, the images demonstrate the application of a machine learning model, specifically using SVM, tested through a web-based information system built with Python's Flask framework. The trained model is saved as a pickle file, which serves as the knowledge base for real-time URL classification.

In these tests, the system successfully classified URLs into the correct categories: image (a) shows a successful classification into the Whitelist category, image (b) correctly identifies a URL in the Porn category, image (c) classifies the URL as Gambling, and image (d) accurately categorizes a URL as Phishing. This demonstrates the model's effectiveness in real-time URL classification.

3.5 Key Findings

In this study, a machine learning-based model was developed to detect negative content, specifically gambling, pornography, and phishing, on websites. The model used three machine learning algorithms: Naïve Bayes, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM), with SVM achieving the highest performance. The model achieved an impressive 90.50% accuracy when tested on a dataset of 2821 active URLs classified as gambling, pornography, phishing, or whitelisted content.

Feature extraction was performed using Term Frequency-Inverse Document Frequency (TF-IDF), incorporating both unigrams and bigrams, thereby significantly enhancing the model's ability to distinguish among content categories. The model was deployed within a Python-based web application to assess the model's effectiveness in real-time, enabling classification of URLs based on user input. This real-time testing confirmed the model's practical viability and potential for deployment in web content moderation systems.

3.6 Discussion

The development and evaluation of this study's negative content detection model present significant advancements in web content moderation. To further understand the contributions of this research, it is essential to compare the key findings with previous works in the literature. The comparison focuses on methodologies, dataset characteristics, classification performance, and real-time implementation, highlighting the similarities and distinctions between this study and others in the field. In particular, the study's approach is compared with that of Chen [41], which incorporates visual and textual content for malicious content detection and explores ensemble classification techniques for identifying malicious webpages [17]. This comparison is summarized in Table 6, which outlines the key differences in findings across these studies.

The comparison indicates that the proposed model achieves competitive performance while using a simpler and more computationally efficient architecture than studies relying on multimodal or ensemble approaches. Although some previous works reported higher accuracy, those methods often required more complex feature engineering or higher computational cost. In contrast, the present study emphasizes practical deployment through lightweight real-time implementation. This practical perspective is consistent with previous work showing that router domain logs can be mined in real time to generate behavioural insights using lightweight and privacy-preserving architectures [42].

Although the current study provides baseline models using SVM, Naïve Bayes, and KNN, further research is required to optimize model performance. Future work will focus on hyperparameter tuning, feature selection,

and evaluating alternative algorithms or ensemble methods to enhance classification accuracy and robustness.

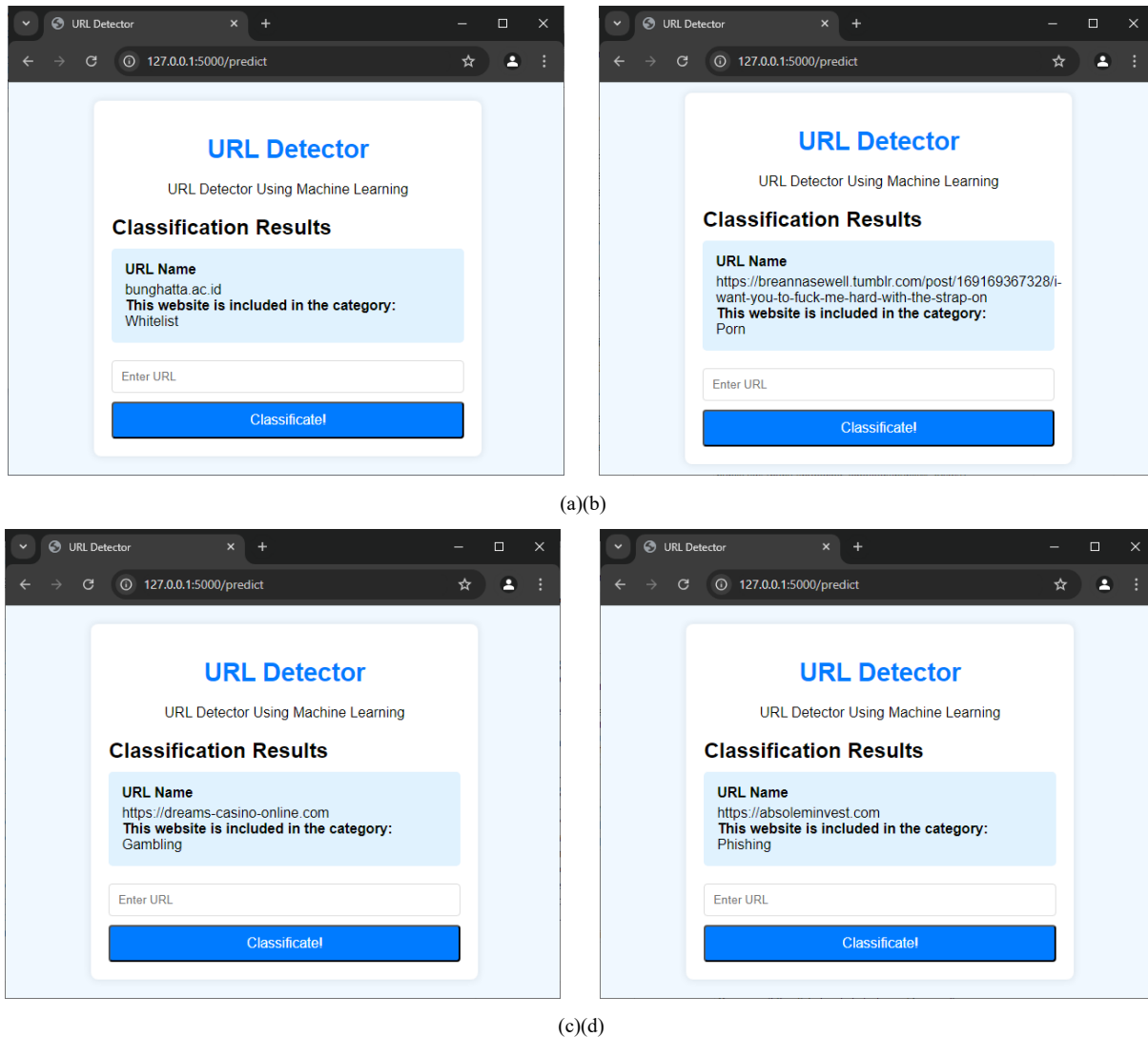


Figure 9. Real-time classification results for (a) whitelist, (b) porn, (c) gambling, and (d) phishing categories

Table 6. Comparison of key findings

Aspect	Research Findings	Article [41]	Article [17]
Detection Approach	Machine learning-based text classification (Naïve Bayes, KNN, SVM)	Combination of visual and textual content using the PG-VTDM method	Ensemble techniques (AdaBoost, Bagging, Random Subspace) for detecting malicious webpages
Dataset	2821 active URLs (multilingual)	The dataset focused on gambling and pornography websites	5000 URLs, 189 features for classifying malicious and benign pages
Model Accuracy	SVM: 90.50%	>99% accuracy (precision, recall, and F-measure)	Best SVM performance in MultiBoost and AdaBoost: 89.15% and 89.05%, respectively
Feature Extraction	TF-IDF (unigrams and bigrams)	Doc2Vec for textual content and Spa-BoVW for visual content	189 features, including dots in URL, WHOIS properties, and DNS
Real-Time Testing	Yes, using a Python-based web application	Yes, applicable to smart routers in homes and schools	No explicit mention of real-time testing, but focuses on ensemble classifiers for webpage classification
Model Delivery	Delivered in pickle (.pkl) format	Not explicitly mentioned	Not explicitly mentioned, focuses on ensemble classifiers

The superior performance of the SVM model can be attributed to its capability to handle high-dimensional feature spaces effectively, particularly when combined with TF-IDF representations. Unlike Naïve Bayes,

which assumes feature independence, and KNN, which is sensitive to noise and data sparsity, SVM constructs an optimal hyperplane that maximizes class separation. Furthermore, the multilingual nature of the dataset

introduces variability in lexical patterns, which SVMs can generalize from more effectively than distance-based approaches. These findings suggest that feature representation and dataset characteristics play a critical role in determining classification performance.

Despite the promising results, several limitations should be acknowledged. First, the dataset was obtained from blacklist repositories and may not fully represent the diversity of real-world websites. Second, some URLs were inactive or inaccessible during the scraping process, reducing the number of usable samples. Third, multilingual content was translated into English, which may introduce minor semantic bias. Future studies should explore larger dynamic datasets and native multilingual language models to improve robustness.

4. Conclusions

This study successfully developed a machine learning-based model for detecting negative website content, utilizing Naïve Bayes, KNN, and SVM algorithms on the multilingual UT1 Blacklist dataset, which includes over ten languages. By focusing on practical algorithms, the research achieves a balance between performance, efficiency, and resource constraints, making the model well-suited for real-time applications.

The feature extraction process, employing TF-IDF with unigrams and bigrams, significantly enhanced classification accuracy, with SVM achieving 90.50% accuracy on 2821 active URLs. Furthermore, implementing a real-time detection system using a Python-based Flask web application addresses a gap in content detection systems. This model provides a reliable solution for mitigating risks posed by harmful online content and provides a robust tool for future content moderation systems.

Additionally, the research outlines future work, including expanding the model for network-based systems to enable dynamic monitoring of website traffic and content detection across various access points. Future work will explore advanced machine learning and deep learning approaches, including transformer-based models, as well as hyperparameter optimization to improve classification performance further. These contributions fill gaps in the existing literature, particularly in the deployment of real-time, multilingual content detection systems, paving the way for future improvements and broader applications.

Acknowledgements

The authors would like to thank Direktur Riset, Teknologi, dan Pengabdian Kepada Masyarakat (DRTPM) of the Direktorat Jenderal Pendidikan Tinggi Kementerian Pendidikan dan Kebudayaan Republik Indonesia for research financial support under Hibah Penelitian Disertasi Doktor with contract no. 102/UN16.19/PT.01.03/2023, and Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) Universitas Andalas for publication support of 2023.

References

- [1] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," *Expert Syst. Appl.*, vol. 153, p. 112986, 2020. doi: [10.1016/j.eswa.2019.112986](https://doi.org/10.1016/j.eswa.2019.112986)
- [2] H. M. Haglan and H. Jasim Ali, "An automatic system for detecting voltage leaks in houses to save people's lives," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 21, no. 3, pp. 1485–1492, 2021. doi: [10.11591/ijeecs.v21.i3.pp1485-1492](https://doi.org/10.11591/ijeecs.v21.i3.pp1485-1492)
- [3] J. E. Reesen et al., "Do better nights lead to better days? Guided internet-based cognitive behavioral therapy for insomnia in people suffering from a range of mental health problems: Protocol of a pragmatic randomized clinical trial," *Contemp. Clin. Trials*, vol. 127, no. December 2022, p. 107122, 2023. doi: <https://doi.org/10.1016/j.cct.2023.107122>
- [4] S. D. Astuti and D. S. F. A. Idola Perdini Putri., "Strategi Komunikasi Program Internet Sehat Dan Aman Kementerian Komunikasi Dan Informatika Republik Indonesia (Studi Kasus Evaluasi Program Incapak Tahun 2015)," *J. Komun.*, vol. 8, no. 1, pp. 25–35, 2016. doi: <https://doi.org/10.24912/jk.v8i1.47>
- [5] H. Chen, Y. Pu, and D. Atkin, "Migration stress, risky Internet uses, and scam victimization: An empirical study among Chinese migrant workers," *Telemat. Informatics*, vol. 83, no. July, p. 102022, 2023. doi: [10.1016/j.tele.2023.102022](https://doi.org/10.1016/j.tele.2023.102022)
- [6] E. D. Fraumenstein and S. Flowerday, "Susceptibility to phishing on social network sites: A personality information processing model," *Comput. Secur.*, vol. 94, p. 101862, 2020. doi: <https://doi.org/10.1016/j.cose.2020.101862>
- [7] B. Sunaryo, J. F. Rusdi, S. Salam, N. A. Abu, M. Sabar, and N. Agustina, "Tracking technology for tourist behavior," in *AIP Conference Proceedings*, 2023, vol. 2691, no. 1, p. 080002. doi: [10.1063/5.0114928](https://doi.org/10.1063/5.0114928)
- [8] C. M. Annur, "Perjudian, Konten Internet Negatif Paling Banyak Diadukan Warga," *Katadata Media Work*, 2022. [Online]. Available: <https://databoks.katadata.co.id/datapublish/2022/08/19/perjudian-konten-internet-negatif-paling-banyak-diadukan-warga>. [Accessed: 19-Jan-2023].
- [9] International Telecommunication Union (ITU), *2021 Measuring digital development: Facts and figures*. 2021. [Online]. Available: <https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2021.pdf>
- [10] D. Kardefelt Winther et al., "Children's exposure to hate messages and violent images online," *UNICEF Innocenti – Glob. Off. Res. Foresight*, pp. 1–8, 2023. [Online]. Available: <https://www.unicef.org/innocenti/documents/childrens-exposure-hate-messages-and-violent-images-online>
- [11] Sumsu, "iGaming Fraud Increased 64 % YoY : Sumsu Shares Alarming Data in 2024 iGaming Fraud Report," *PR Newswire*, 2024. [Online]. Available: <https://www.prnewswire.com/news-releases/igaming-fraud-increased-64-yoy-sumsub-shares-alarming-data-in-2024-igaming-fraud-report-302151125.html>. [Accessed: 02-Oct-2025].
- [12] C. Hooper, "2025 iGaming fraud prevention guide : trends , types , and regulations," *Veriff*, 2025. [Online]. Available: <https://www.veriff.com/fraud/learn/igamingfraud>. [Accessed: 02-Oct-2025].
- [13] "What is online gambling fraud, and how can it be prevented?," *ComplyAdvantage*, 2024. [Online]. Available: <https://complyadvantage.com/insights/online-gambling-fraud/>. [Accessed: 02-Oct-2025].
- [14] "Online Gambling Global Market Report 2025," *The Business Research Company*, 2025. [Online]. Available: <https://www.thebusinessresearchcompany.com/report/online-gambling-global-market-report>. [Accessed: 02-Oct-2025].
- [15] A. Gregory, "Gambling poses huge global threat to public health, experts warn," *The Guardian*, 2024. [Online]. Available: <https://www.theguardian.com/society/2024/oct/24/gambling-poses-huge-global-threat-to-public-health-experts-warn-lancet-commission>. [Accessed: 02-Oct-2025].
- [16] A. Amalia, D. Gunawan, M. S. Lydia, and Wesley, "The Identification of Negative Content in Websites by Using

- Machine Learning Approaches,” *5th Int. Conf. Comput. Eng. Des. ICCED 2019*, 2019. doi: [10.1109/ICCED46541.2019.9161105](https://doi.org/10.1109/ICCED46541.2019.9161105)
- [17] A. Subasi, M. Balfagih, Z. Balfagih, and K. Alfawwaz, “A Comparative Evaluation of Ensemble Classifiers for Malicious Webpage Detection,” *Procedia Comput. Sci.*, vol. 194, pp. 272–279, 2021. doi: <https://doi.org/10.1016/j.procs.2021.10.082>
- [18] P. K. Mvula, P. Branco, G. V. Jourdan, and H. L. Viktor, “COVID-19 malicious domain names classification,” *Expert Syst. Appl.*, vol. 204, no. May, 2022. doi: [10.1016/j.eswa.2022.117553](https://doi.org/10.1016/j.eswa.2022.117553)
- [19] N. Izzah, I. Budi, and S. Louvan, “Classification of Pornographic Content on Twitter Using Support Vector Machine and Naive Bayes,” *2018 4th Int. Conf. Comput. Technol. Appl.*, pp. 156–160, 2018. doi: [10.1109/CATA.2018.8398674](https://doi.org/10.1109/CATA.2018.8398674)
- [20] M. I. Rusydi et al., “Electrooculography signal as alternative method to operate wheelchair based on SVM classifier,” in *AIP Conference Proceedings*, 2024, vol. 2891, no. 1, p. 020013. doi: <https://doi.org/10.1063/5.0200941>
- [21] S. Bahri, K. Saddami, F. Arnia, and K. Muchtar, “Perbandingan Kinerja Support Vector Machine (SVM) Dalam Mengenali Wajah Menggunakan SURF DAN GLCM,” *J. Nas. Tek. Elektro*, vol. 8, no. 2, p. 65, 2019. doi: [10.25077/jnte.v8n2.620.2019](https://doi.org/10.25077/jnte.v8n2.620.2019)
- [22] B. B. Gupta, K. Yadav, I. Razzak, K. Psannis, A. Castiglione, and X. Chang, “A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment,” *Comput. Commun.*, vol. 175, no. November 2020, pp. 47–57, 2021. doi: [10.1016/j.comcom.2021.04.023](https://doi.org/10.1016/j.comcom.2021.04.023)
- [23] Fabrice Prigent, “Blacklists UT1,” 2023. [Online]. Available: http://dsi.ut-capitole.fr/blacklists/index_en.php. [Accessed: 11-Aug-2023].
- [24] U. T. 1 Capitole, “UT1-blacklist,” 2023. [Online]. Available: <https://github.com/olbat/ut1-blacklists>. [Accessed: 11-Aug-2023].
- [25] M. Safaei Pour, C. Nader, K. Friday, and E. Bou-Harb, “A Comprehensive Survey of Recent Internet Measurement Techniques for Cyber Security,” *Comput. Secur.*, vol. 128, p. 103123, 2023. doi: <https://doi.org/10.1016/j.cose.2023.103123>
- [26] N. R. Njeri, “Data Preparation For Machine Learning Modelling,” *Int. J. Comput. Appl. Technol. Res.*, vol. 11, no. 06, pp. 231–235, Jun. 2022. doi: [10.7753/IJCATR1106.1008](https://doi.org/10.7753/IJCATR1106.1008)
- [27] M. R. Bingeri, S. Eswaran, and P. Honnavalli, “Malicious Network Traffic Detection in Internet of Things Using Machine Learning,” in *Proceedings of Data Analytics and Management*, A. Khanna, Z. Polkowski, and O. Castillo, Eds. Singapore: Springer Nature Singapore, 2023, pp. 707–718. doi: https://doi.org/10.1007/978-981-19-7615-5_58
- [28] K. Munawaroh and A. Alamsyah, “Performance Comparison of SVM, Naïve Bayes, and KNN Algorithms for Analysis of Public Opinion Sentiment Against COVID-19 Vaccination on Twitter,” *J. Adv. Inf. Syst. Technol.*, vol. 4, no. 2, pp. 113–125, Mar. 2023. doi: [10.15294/jaist.v4i2.59493](https://doi.org/10.15294/jaist.v4i2.59493)
- [29] S. F. Ahmed et al., “Deep learning modelling techniques: current progress, applications, advantages, and challenges,” *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 13521–13617, Nov. 2023. doi: <https://doi.org/10.1007/s10462-023-10466-8>
- [30] S. K. Tripathy, R. Singh, R. Srivastava, A. K. Bhoi, and S. K. Satapathy, “Datasets BT - Advances in Human Activity Detection and Recognition (HADR) Systems,” *Synthesis Lectures on Computer Science*, Eds. Cham: Springer Nature Switzerland, 2024, pp. 33–57. doi: [10.1007/978-3-031-51660-3](https://doi.org/10.1007/978-3-031-51660-3)
- [31] J. F. Rusdi et al., “Dataset smartphone usage of international tourist behavior,” *Data Br.*, vol. 27, p. 104610, Dec. 2019. doi: [10.1016/S1389-1286\(01\)00256-0](https://doi.org/10.1016/S1389-1286(01)00256-0)
- [32] K. H. Cheung and J. Miši, “On virtual private networks security design issues,” *Comput. Networks*, vol. 38, no. 2, pp. 165–179, 2002. doi: [10.1016/S1389-1286\(01\)00256-0](https://doi.org/10.1016/S1389-1286(01)00256-0)
- [33] W. W. Daniel, *Biostatistics: A Foundation for Analysis in the Health Sciences.*, vol. 44, no. 1. 1988.
- [34] T. Kudo, T. Yamamoto, and T. Watanabe, “Three-Step Master Data Creation Method from Big Data: Scraping, Semi-Structuring, and Extraction,” *Procedia Comput. Sci.*, vol. 207, no. Kes, pp. 360–369, 2022. doi: <https://doi.org/10.1016/j.procs.2022.09.070>
- [35] Y. A. Mahmood and B. Mahmood, “A Web Scraper for Data Mining Purposes,” *Sistemasi*, vol. 13, no. 3, p. 1243, 2024. doi: [10.1016/j.dib.2024.110906](https://doi.org/10.1016/j.dib.2024.110906)
- [36] H. N. Prasetyo, R. Sarno, D. R. Wijaya, R. Budiraharjo, I. Waspada, and K. R. Sungkono, “Data of dwelling time process at container terminal : Multi-perspective based dataset,” *Data Br.*, no. xxxx, p. 110906, 2024. doi: [10.1016/j.dib.2024.110906](https://doi.org/10.1016/j.dib.2024.110906)
- [37] B. Sunaryo, M. I. Rusydi, A. Hazmi, and M. Sasaki, “A Systematic Literature Review of Automation Quality of Service in Computer Networks: Research Trends, Datasets, and Methods,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 2, pp. 353–366, Mar. 2023. doi: [10.29207/resti.v7i2.4810](https://doi.org/10.29207/resti.v7i2.4810)
- [38] Y. Setiawan, N. Ulfa Maulidevi, and K. Surendro, “The Optimization of n-Gram Feature Extraction Based on Term Occurrence for Cyberbullying Classification,” *Data Sci. J.*, vol. 23, no. 1, pp. 1–21, May 2024. doi: [10.5334/dsj-2024-031](https://doi.org/10.5334/dsj-2024-031)
- [39] J. Zhou, Z. Ye, S. Zhang, Z. Geng, N. Han, and T. Yang, “Investigating response behavior through TF-IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data,” *Heliyon*, vol. 10, no. 16, p. e35945, Aug. 2024. doi: <https://doi.org/10.1016/j.heliyon.2024.e35945>
- [40] M. Ahmadi, M. Khajavi, A. Varmaghani, A. Ala, K. Danesh, and D. Javaheri, “Leveraging large language models for cybersecurity: enhancing SMS spam detection with robust and context-aware text classification,” *Cyber-Physical Syst.*, pp. 1–27, Sep. 2025. doi: [10.48550/arXiv.2502.11014](https://doi.org/10.48550/arXiv.2502.11014)
- [41] Y. Chen, R. Zheng, A. Zhou, S. Liao, and L. Liu, “Automatic Detection of Pornographic and Gambling Websites Based on Visual and Textual Content Using a Decision Mechanism,” *Sensors*, vol. 20, no. 14, p. 3989, Jul. 2020. doi: <https://doi.org/10.3390/s20143989>
- [42] B. Sunaryo, M. I. Rusydi, A. Hazmi, and R. A. Ramllee, “Real-Time Mining of Router Domain Logs for Temporal and Behavioural Analysis in Smart Network Environments,” in *2025 5th International Conference of Science and Information Technology in Smart Administration (ICSINTESA)*, 2025, pp. 137–142. doi: [10.1109/ICSINTESA68165.2025.11413661](https://doi.org/10.1109/ICSINTESA68165.2025.11413661)