

CHAPTER 2 - LITERATURE REVIEW

This chapter will present a comprehensive literature review as a theoretical foundation for research on optimizing the effectiveness of a directed machine learning approach supervised learning to predict the compliance of ISPs in Indonesia towards negative content data synchronization. This literature review aims to provide an in-depth understanding of the issues and reveal research gaps that will be addressed through this study, so as to make a significant contribution to predict the compliance of ISPs in Indonesia towards negative content data synchronization.

2.1. Ministry of Communication Informatic

In accordance with Law No. 39/2008 on Ministries of State, the Ministry of Communication and Informatics is an institution of the Government of the Republic of Indonesia in charge of the scope of activities as stated in the 1945 Constitution of the Republic of Indonesia, namely information and communication. The Ministry of Communication and Informatics was previously named “Department of Information” from 1945-1999. Subsequently, from 2001 to 2005, it changed its name to “Ministry of State Communication and Information” (2001-2005), and “Department of Communication and Information” (Depkominfo) (2005-2009). The Ministry of Communication and Information has the task of organizing government affairs in the field of communication and informatics to assist the President in organizing the state government. The Ministry of Communication and Information is led by a Minister of Communication and Information Technology (Menkominfo) (Direktorat Jenderal Aplikasi Informatika, 2021).

The Directorate General of Informatics Applications (Aptika), which was originally the Directorate General of Telematics Applications (Aptel), was established in order to develop effective Information and Communication Technology (ICT) in Indonesia. ICT cannot be separated from the 3C elements, namely Communication, Computer, and Content, which were then initiated as the Kemkominfo logo which is a composition of three letters C in blue gradations. The Directorate General of Informatics Applications was established in accordance with the mandate of the Minister of Communication and

Informatics Regulation Number 12 of 2021 concerning the Organization and Work Procedures of the Ministry of Communication and Informatics. Aptika has the task of organizing the formulation and implementation of policies in the field of informatics application management.

2.2. Existing Regulations Related to Domain and IP Blocking in Indonesia

While the regulations mentioned previously provide a framework for regulating online content, here are some specific regulations and decrees relevant to the blocking of domains and IP addresses in Indonesia. The key regulations are as follows:

1. Law Number 11 of 2008 concerning Electronic Information and Transactions (UU ITE) as amended by Law Number 19 of 2016 concerning Amendments to Law Number 11 of 2008 concerning Electronic Information and Transactions in article 40 paragraph (2): A. The Government is obliged to prevent the dissemination and use of Electronic Information and / or Electronic Documents that have prohibited content in accordance with the provisions of laws and regulations; B. In preventing the dissemination and use of Electronic Information and / or Electronic Documents that have prohibited content, the Government is authorized to cut off access and / or order the Electronic System Operator to cut off access to Electronic Information and / or Electronic Documents that have unlawful content.
2. PP No. 71 of 2019 concerning the Implementation of Electronic Systems and Transactions in article 96, Termination of Access is carried out to Electronic Information and / or Electronic Documents with classification: a. violating laws and regulations; b. disturbing the public and disturbing public order; and c. informing how or providing access to Electronic Information and / or Electronic Documents that have prohibited content in accordance with the provisions of laws and regulations.
3. Regulation of the Minister of Communication and Information No. 5 of 2020 concerning Private Scope Electronic System Operator and its amendments in article 14 paragraph (1), Requests for Termination of Access to Electronic Information and / or Electronic Documents that are prohibited can be submitted by: a. the public; b. Ministry or Institution; c. Law Enforcement Officials; and /

- or d. Judicial institutions. Judicial institutions. (Menteri Komunikasi Dan Informatika Republik Indonesia, 2020)
4. Minister of Communication and Informatics Regulation No. 10 of 2021 concerning the Second Amendment to Minister of Communication and Informatics Regulation No. 13 of 2019 in article 18: 1) Telecommunication Service Providers are prohibited from providing services that have content contrary to the provisions of laws and regulations; 2) In the event that the Telecommunication Service Provider does not provide content directly, the Telecommunication Service Provider must inform the prohibition of content as referred to in paragraph (1) to partners or parties who cooperate in providing content; 3) In the event of a complaint of content that violates the provisions of laws and regulations as referred to in paragraph (1), the Telecommunication Service Provider must block the service in question. (Menteri Komunikasi Dan Informatika Republik Indonesia, 2021)

These regulatory measures reflect Indonesia's commitment to creating a secure financial environment where consumers can trust that their assets and personal data are well-protected. The integration of these regulations into the operations of financial institutions is crucial for building a resilient financial sector capable of withstanding the evolving landscape of cyber threats and maintaining consumer confidence.

2.3. APJII

APJII is an association formed on May 15 1996 for the development of internet networks in Indonesia. Its members are spread throughout Indonesia and act as government partners in building national and international information and communication facilities. APJII also has a role in assisting arbitration efforts, namely mediating, reconciling and resolving disputes between its members. APJII has several main services, including:

- APJII Membership : APJII accommodates internet service provider (ISP) business actors in Indonesia.
- IDNIC (Indonesia Network Information Center) is the National Internet Registry that manages IP addresses for Indonesia.

- IIX (Indonesia Internet Exchange) is an internet traffic exchange point in Indonesia that allows interconnection between ISPs. The formation of IIX is one of APJII's efforts to facilitate internet availability in the regions.

APJII regularly releases reports on Indonesia's internet profile, such as "Indonesia Internet Profile 2022". This report provides a comprehensive picture of internet development in Indonesia, including statistical data on users, accessibility and internet usage in various sectors. To carry out its functions, APJII has an organizational structure consisting of a National Conference, a Board of Protectors/Guardians, a Management Board, a Daily Executive Body, and a Working Group (Pokja). This organizational structure ensures that APJII can operate effectively and efficiently in achieving its goals. Apart from that, APJII also has representatives in various regions in Indonesia, such as APJII East Java, which is tasked with accommodating ISPs in that region. The presence of representatives in this area shows APJII's commitment to advancing the internet throughout Indonesia.

2.4. ISP

ISP is a company that provides internet access services for customers¹. ISPs in Indonesia offer various types of services with various technologies, so customers can choose services that suit their needs and budget. APJII is an organization that oversees ISPs in Indonesia¹⁶. APJII plays a role in regulating and influencing ISPs in Indonesia through various means (APJII, 2024), such as :

- Providing IIX (Indonesia Internet Exchange) connection services. IIX is an important platform that enables interconnection between ISPs, thereby increasing the efficiency and quality of internet services in Indonesia.
- Manage IP Address and AS Number allocation via APJII NIR. APJII NIR ensures that the allocation of IP addresses and AS Numbers is carried out in an orderly manner and in accordance with international standards.

APJII currently has more than 1000 ISP members in Indonesia. In 2024, the number of internet users in Indonesia will reach 221 million people out of a total population of 278

million people, with an internet penetration rate of 79.5 persen. This table 2.1 shows a significant increase in recent years.

Table 2.1 Number of Members and Internet Penetration

Year	Internet Penetration Rate
2018	64,8%
2020	73,7%
2022	77,01%
2023	78,19%
2024	79,5%

2.5. Domain Name Server (DNS)

DNS is short for Domain Name System (DNS server), which is the name of the DNS server. a database system that is useful for meeting the needs of computers, services or resources connected to the internet or computer networks or resources connected to the internet or private computer networks. private computer network. DNS allows the name of a host on a computer network or the internet to be translated into an IP address, and vice versa. Address, and vice versa (Ichise et al., 2023). This system works like a phone book for the internet. Where the domain name represents the name of the website and the IP Address represents the website's location on the internet. The basic principle of how DNS works is to match the URL component name with the IP Address component. Each URL and IP Address has parts that explain each other. DNS servers are an important part of the process. Domain Name Server is a system in charge of translating an IP Address to its URL or vice versa, DNS helps users find the intended website or address according to input from the user, be it IP Address or URL (Knerler et al., 2022) It can be seen in figure 2.1.

In the picture above, the difference is that the library code starts from the front. Meanwhile, the code on DNS is sorted from the back. The following is a complete explanation :

1. Root-Level Domain is the highest part of the DNS hierarchy. It usually takes the form of a dot (.) at the very end of a URL.

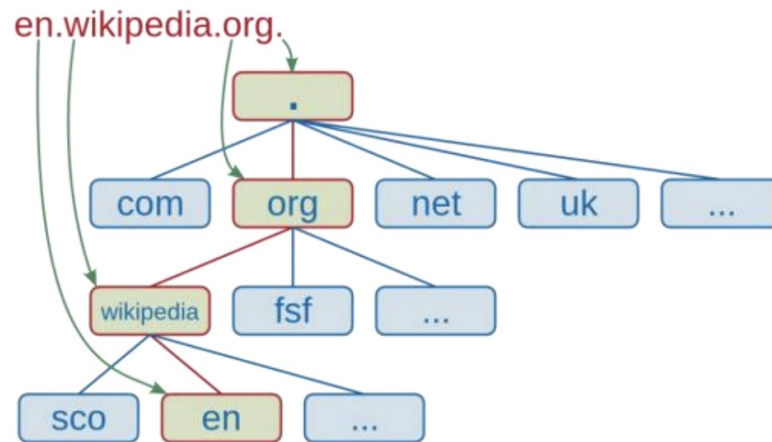


Figure 2.1 Model Tree of Domain

2. Top Level Domain is the extension that sits in front of the root-level domain. There are two types of TLDs that are commonly used. They are Generic Top-Level Domain (GTLD) and Country Code Top Level Domain (CCLTD). The GTLD usually describes the institutional nature of the web owner. Let's say, a website for commercial purposes usually has a .COM extension. Then, .EDU for educational institutions and .GO.ID for government agencies.

2.6. Response Policy Zone (RPZ)

A more granular and widely used method for content filtering operates at the Domain Name System (DNS) layer. DNS filtering allows an ISP to control access to websites based on their domain names rather than their IP addresses. One of the most powerful mechanisms for this is the use of Response Policy Zones (RPZ). RPZ is a feature within DNS servers that enables an administrator to override the normal DNS resolution process according to a set of predefined policies.

Many researches have been done in the literature for detecting and blocking botnet-based cyber attacks. In the authors devised a detection system for worm infected computers sending mass-mailing by analyzing DNS traffic using the Bayesian method. The DNS traffic was obtained in about two hours from ISPs (Internet Service

Providers). In particular, the A record and the corresponding MX record were analyzed using the Bayesian method .

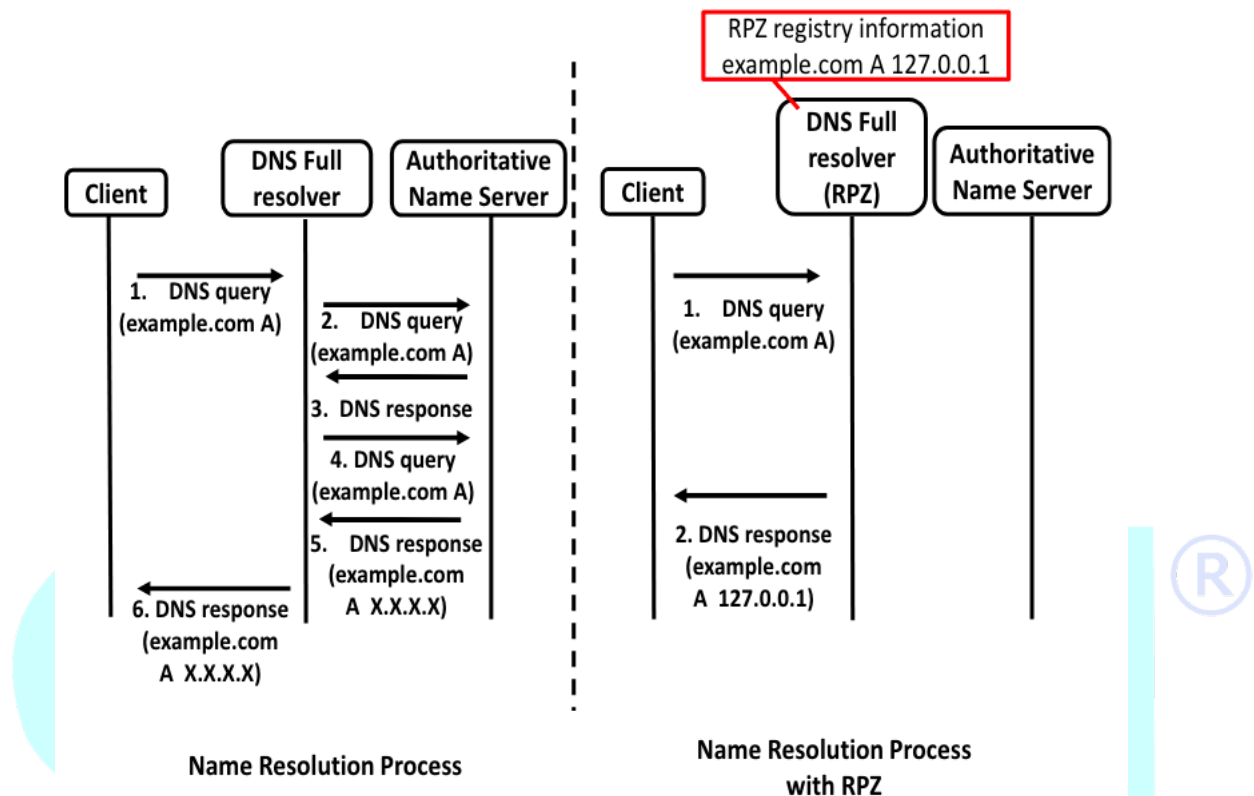


Figure 2.2 Overview of DNS RPZ

In the figure 2.2 DNS RPZ is a special authoritative zone in a DNS server for policy-based control of the target domain name and it is a special feature of BIND. Figure 4 shows a brief procedure of domain name resolution involving DNS RPZ. In a normal name resolution process without involving DNS RPZ, the client sends a DNS query to the DNS full-service resolver first. Then the DNS full-service resolver queries to the authoritative DNS servers iteratively and replies the answer to the client. Finally, the client receives the DNS response from the DNS full-service resolver.

2.7. DNS Trust Next Generation (TNG)

According to regulations from the Ministry of Communication and Information Technology (Kominfo) in 2024, the Directorate of PPI Control and the Directorate of Aptika Control are conducting trials of new content blocking using two methods, namely the Remotely Triggered Black Hole (RTBH) method for IP address blocking

and DNS Trust Next Generation (DNS Trust-NG) for domain address blocking. (Direktorat Pengendalian PPI, 2024) DNS Trust-NG is the DNS Trust New Generation, an enterprise DNS server provided by Kominfo as a solution to :

1. Implementation challenges of the trust+ filter on the ISP side due to various reasons, as evidenced by many ISPs still not implementing comprehensive filtering on the internet network distributed to the public. The presence of Trust-NG simplifies ISP implementation of the trust+ filter.
2. The growing size of the Trust+ database, which is too heavy to run via RPZ; Trust-NG is much lighter and performs better.
3. The need to monitor internet service providers' compliance with their obligation to implement filtering based on the Trust+ database issued by the government, which could not be monitored on the RPZ system previously.

Since its inception, DNS Trust NG has carried the Trust+ database, which syncs every minute to the Master to check for database changes, such as domain additions or reductions, so that it can relatively instantly follow the latest changes. DNS Trust NG is designed to perform at its best (maximum performance), be easier to implement (plug-and-play), simpler to operate, compatible with various types of PCs/servers, lighter, and faster in performance compared to other systems (true performance). DNS Trust NG is designed to perform at its best (maximum performance).

2.8. Remotely Triggered Black Hole (RTBH) Filtering

Remotely Triggered Black Hole (RTBH) filtering is a powerful network security technique used to drop undesirable traffic at the edge of a network, preventing it from consuming internal resources. Although its primary utility lies in the mitigation of large-scale Distributed Denial-of-Service (DDoS) attacks, the mechanism itself possesses direct applicability to content blocking.

The process is typically initiated by a central "trigger" device within the network operations center (NOC). When a target IP address needs to be blocked, the trigger device injects a routing update into the network's Internal Border Gateway Protocol (iBGP). This update advertises the target IP address with a next-hop that points to a pre-configured discard route on the network's edge routers. Upon receiving this update, all edge routers learn this new, specific route. Subsequently, any inbound packet destined

for the blackholed IP address is immediately forwarded to the null interface and dropped, effectively blocking all communication with that server.

In the context of Indonesian content filtering, RTBH serves as a foundational, albeit coarse, tool. When Kominfo issues a directive to block a website or service, the ISP can identify the IP address of the hosting servers and use RTBH to null-route them. This method is effective and rapid but is also a "blunt instrument". It blocks all forms of traffic to the IP address indiscriminately, which can lead to over-blocking if the server hosts multiple websites or services, only one of which is on the blacklist.

2.9. Machine Learning

Machine learning (ML) is a branch of artificial intelligence (AI) which is also a branch of statistics which focuses on developing algorithms that allow computers to learn from data and perform tasks without explicit instructions. ML involves the study and construction of algorithms that can learn from and make predictions on data. Arthur Samuel, a pioneer in the field of AI, defined machine learning as "a field of study that gives computers the ability to learn without being explicitly programmed". Modern machine learning has two main goals :

- Data classification for group data into different categories
- Prediction about future events based on historical data

For example, machine learning algorithms can be used to classify emails as spam or not spam, or to predict future stock prices. ML allows computers to “think” and act like humans 6, enhancing their abilities in perception, cognition, and action through experience. Machine learning can be categorized into several types based on how the algorithm learns and the type of data used. The four main types of machine learning are :

1. Supervised Learning

As a subset of machine learning, supervised learning involves training algorithms on a dataset where each input example is associated with a corresponding correct output label.

2. Unsupervised Learning

Unsupervised learning is a type of machine learning where algorithms are trained using unlabeled data

Comparison table of unsupervised learning and supervised learning :

Table 2.2 Comparison table of unsupervised learning and supervised learning

No	Parameter	Unsupervised Learning	Supervised Learning
1	Input Data	The algorithm is used on unlabeled data	The algorithm is trained using labeled data
2	Computational Complexity	Very computationally complex	Simpler
3	Accuracy	Very accurate and trustworthy	Less accurate but still reliable

2.10. Supervised Machine Learning

Supervised learning is a type of machine learning where algorithms are trained using labeled data. Training data consists of the desired input and output pairs. The algorithm learns from this data to map the input to the correct output, and then uses this mapping to predict the output for new data that it has never seen before. Steps in supervised learning :

- Identify training data types
- The data must be similar to the input data that will be processed by the model
- Collect and label training data
- Training data must be free from bias.
- Create three groups of data training data, validation data, and test data.
- Choose a machine learning algorithm to create a model.
- Train the model with training data.

- Model evaluation with validation data and test data.

Supervised learning is suitable for prediction tasks when labeled data is available. Predict the creditworthiness of potential borrowers based on their credit history and synchronize data ISP Compliance.

2.11. Random Forest

Random Forest is a machine learning algorithm that falls into the ensemble learning category. Ensemble learning combines multiple machine learning models to improve prediction accuracy and stability. Random Forest builds many decision trees and combines the results to get more accurate and robust predictions.

Random Forest builds a large number of decision trees using different data subsets and randomly selected features. Training each decision tree: Each decision tree is trained on a different subset of data. Combining prediction results: To get the final prediction, Random Forest combines the prediction results from all decision trees. In the case of classification, the final prediction is determined based on majority voting from all trees. In the case of regression, the final prediction is the average of the predictions of all trees.

In the context of predicting Internet Service Provider (ISP) compliance with negative content data synchronization, the Random Forest algorithm provides a significant advantage due to its ability to manage complex interactions between variables (Utamima et al., 2023). For instance, when predicting compliance, factors such as user behavior, content type, and historical compliance records can interact in non-linear ways. The Random Forest's capacity to capture these interactions makes it a suitable choice for this research.

Statistically, Random Forest has been shown to outperform many traditional machine learning algorithms in various applications, particularly in classification problems. According to a study by Caruana and Niculescu-Mizil (2006), Random Forest achieved an accuracy rate of 93.5% in predicting the presence of diseases in a medical dataset, compared to 89.6% for Logistic Regression and 91.4% for Support Vector Machines

(SVM). Such performance metrics indicate the potential of Random Forest in tasks requiring high predictive accuracy, such as assessing ISP compliance.

Furthermore, the interpretability of Random Forest models, while not as straightforward as single decision trees, can still be enhanced through techniques like variable importance measures. These measures help identify which features contribute most significantly to the predictions, providing insights into the underlying factors influencing ISP compliance (Liaw and Wiener, 2002). This interpretability is crucial for stakeholders in the ISP sector, as it allows them to understand compliance dynamics better and implement informed strategies.

2.12. Logistic Regression

Logistic Regression is a machine learning algorithm used to predict the probability of an event. This algorithm falls into the supervised learning category and is generally used for binary classification problems, where the output has only two possibilities, such as "yes" or "no".

Logistic Regression uses a logistic function to model the relationship between input variables and output probabilities. The logistic function produces a value between 0 and 1, which represents the probability of an event.

A statistical model that is highly effective for binary classification. It is chosen for its simplicity, computational efficiency, and high degree of interpretability, as the model's coefficients directly indicate the influence of each feature on the outcome. It serves as an excellent and powerful baseline for comparison.

2.13. Support Vector Machines (SVM)

SVM is a supervised learning algorithm used for classification and regression. This algorithm looks for a hyperplane that separates data into two categories with the largest margin. SVM is effective for high-dimensional data, but can be complex for large datasets. A powerful algorithm that works by finding the optimal hyperplane that best separates the data points of different classes in a high-dimensional space. SVMs are

selected for their effectiveness in creating clear separation margins and their ability to handle non-linear relationships through the use of kernels.

2.14. Linear Regression

Linear regression is a supervised learning algorithm designed for the prediction of numerical values, leveraging the discernible linear relationship between independent and dependent variables. This algorithm models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. For instance, in predicting Internet Service Provider (ISP) compliance with negative content data synchronization, Linear Regression can provide insights into how different factors influence compliance rates. The significance of Linear Regression lies in its ability to predict outcomes based on historical data. According to a study by (Moreira and Fachine, 2018), Linear Regression is a widely adopted statistical method in diverse academic domains, notably economics, biology, and the social sciences, attributed to its robust capability in quantifying relationships among variables. For example, in the context of ISP compliance, factors such as user behavior, content type, and regulatory measures can be quantitatively analyzed to predict compliance levels. This predictive capability is crucial for ISPs to align their services with legal requirements and societal expectations.

2.15. KNN

KNN is a supervised learning algorithm used for classification and regression. This algorithm classifies new data based on its proximity to the training data. KNN is simple and easy to implement, but can be slow for large datasets.

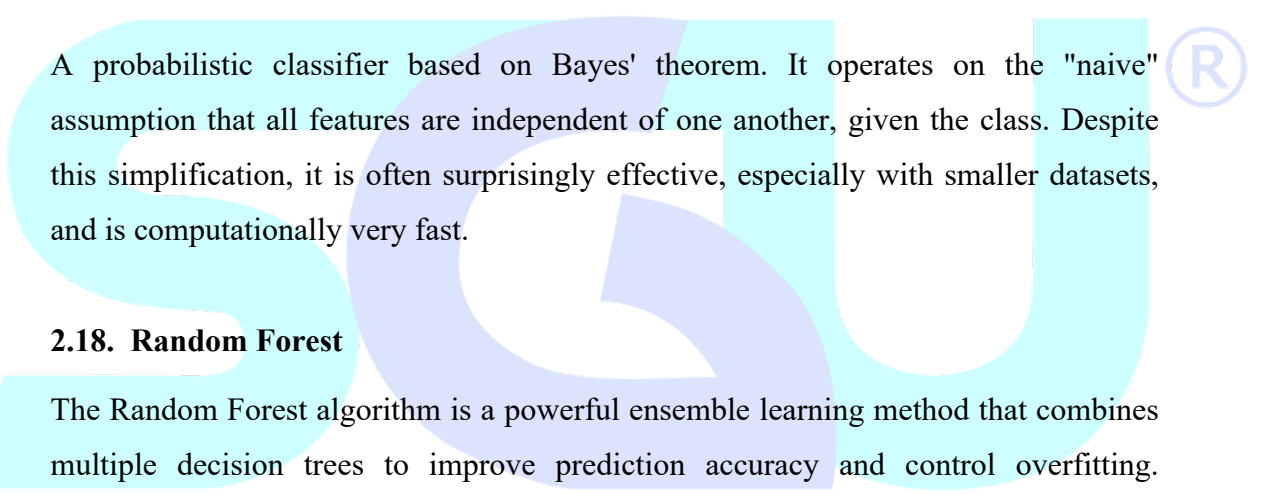
2.16. Decision Trees

Decision tree is a supervised learning algorithm used for classification and regression. This algorithm builds a tree-like model where each node represents a feature, each branch represents a decision rule, and each leaf represents an outcome. Decision trees are easy to read and interpret, but can be unstable and less effective for continuous variables. A non-parametric model that predicts the target variable by learning simple

decision rules inferred from the data features. It is included primarily for its unparalleled interpretability; the model can be visualized as a flowchart, making its decision-making logic transparent and easily communicable to non-technical stakeholders.

2.17. Naive Bayes

Naive Bayes is a simple and effective machine learning algorithm for classification. This algorithm is based on Bayes' theorem, which calculates the probability of an event based on prior knowledge. Naive Bayes assumes that each feature is independent of every other feature, which is called "naive" because this assumption is rarely met in the real world. Naive Bayes calculates the probability that data falls into a certain class based on the probability of each feature in that class.

A probabilistic classifier based on Bayes' theorem. It operates on the "naive" assumption that all features are independent of one another, given the class. Despite this simplification, it is often surprisingly effective, especially with smaller datasets, and is computationally very fast. 

2.18. Random Forest

The Random Forest algorithm is a powerful ensemble learning method that combines multiple decision trees to improve prediction accuracy and control overfitting. Introduced by Breiman in 2001, Random Forest has gained popularity due to its robustness and ability to handle large datasets with high dimensionality (Febrianto and Soediantono, 2022). This algorithm operates on the principle of bagging, where multiple subsets of the training data are generated, and individual decision trees are trained on these subsets. The final prediction is obtained by aggregating the predictions from all individual trees, typically through majority voting for classification tasks or averaging for regression tasks (Liaw and Wiener, 2002).

In the context of predicting Internet Service Provider (ISP) compliance with negative content data synchronization, the Random Forest algorithm provides a significant advantage due to its ability to manage complex interactions between variables (Utamima et al., 2023). For instance, when predicting compliance, factors such as user

behavior, content type, and historical compliance records can interact in non-linear ways. The Random Forest's capacity to capture these interactions makes it a suitable choice for this research.

2.19. Successful Metrics

In the quest to achieve operational excellence within the Response Policy Zone (RPZ), the selection and implementation of 'good metrics' emerges as a critical element. Good metrics are not just numbers; they are windows through which to view and evaluate the actual performance of the RPZ. Successful metrics are those that meet criteria that are specific, measurable, assignable, realistic, and time-related (SMART methods) or have certain quality criteria, including precision, measurability, meaningfulness, and usefulness (Forsberg and Frantti, 2023). These successful metrics are designed to measure the performance of RPZs and address limitations observed in the existing landscape of RPZ-related metrics. Table 2.3 explain the definition for each component.

Table 2.3 S.M.A.R.T Goals

No	Component	Explanation	Example
1	Specific	Goals should be clear and precise.	"Reduce malware detection and response time by 30% within the next quarter."
2	Measurable	Metrics should be quantifiable to track progress.	Data on the number of incidents resolved, time to detect and respond to threats, or reduction in false positives.
3	Achievable	Goals must be realistic given current resources.	Aim to "improve threat detection accuracy by 20%" rather than "eliminate all cyber threats."
4	Relevant	Goals need to align with broader security objectives.	If susceptible to ransomware, a goal might be "enhance defenses against ransomware attacks."

5	Time-bound	Goals should have a clear timeline.	“Deploy a new intrusion detection system by the end of Q3.”
---	------------	-------------------------------------	---

2.19.1 Quality of Metrics

There is a paper from (Laayati et al., 2022) mention a good metrics should following this quality :

1. Relevance and Meaningfulness

Metrics should measure factors that are important and relevant to the specific objectives and context of the organization or operation. They should not only focus on what is crucial but also make sense to the people who will use them. For example, in a RPZ, metrics might include the number of detected threats, the time taken to respond, and the effectiveness of response strategies.

2. Cost-Effectiveness

The collection and processing of metrics should not be so costly or resource-intensive that they detract from other necessary operations or stages of measurement. The resources allocated to developing and maintaining metrics should be justified by the value they bring to the organization.

3. Timeliness and Frequency

The timing and regularity of measurements within a Response Policy Zone (RPZ) must align with how quickly the security landscape changes. This alignment ensures that the metrics remain relevant and effective. If there's too much delay in measuring and updating these metrics, they could become obsolete and fail to serve their intended purpose of guiding security operations. Additionally, it's crucial for these metrics to allow for tracking and analyzing trends over time, enabling RPZ teams to adapt to evolving threats and improve their defensive strategies.

4. Objective and Quantifiable

Effective metrics should be objective and quantifiable, meaning they are based on specific, reliable numerical data rather than subjective qualitative evaluations, which can be prone to bias. These metrics should also be presented

in clear and straightforward units of measurement that are easy to understand and interpret.

Another paper from (Ichise et al., 2023) mention a metrics should have this quality :

1. Correctness

- Granularity : The metric should offer sufficient detail to link it to a specific function or team within a RPZ.
- Completeness : The metric should fully achieve the objective defined in its documentation. If it cannot do so alone, it should be paired with another metric.
- Objective and unbiased : The metric's results should remain unaffected by the activities of the person setting it up, with bias minimized to acceptable levels as specified in the metric description.

2. Measurability

- Availability : Data used to create the metric must be automatically accessible in a reliable and consistent format.
- Reproducibility : The metric should be replicable by different individuals across various organizations, provided they have access to the same data.

3. Meaningfulness

- Impactful : The metric should influence daily activities and effectively demonstrate the progress of development efforts
- Clarity : The metric's interpretation should be clear and consistent throughout its lifecycle.
- Comparability : The metric's results should be comparable across different RPZs, even between different organizations.

4. Usability

- Portability : The metric should be applicable to various RPZs, regardless of their size, structure, service model, or parent organization.
- Controllability : The team being measured should be able to maintain the metric value within expected ranges.
- Scalability : The metric should perform consistently regardless of the volume of measurements.

2.20. Related Works

A study by (Ichise et al., 2023) developed a comprehensive framework aimed at improving and evaluating the performance of Response Policy Zones (RPZs). The framework was constructed through a detailed literature review and incorporated a point-based metrics evaluation system. This study provided foundational insights into the critical components and processes that contribute to an effective RPZ, highlighting areas for potential improvement and benchmarking. The metrics discussed in this study include processor core utilization, RAM memory usage, disk throughput, log collectors' performance, log processors' efficiency, skilled analyst count, and average time to complete one offense.

Varshney and Shah, in their 2021 paper (Varshney and Shah, 2021), proposed a framework specifically designed to measure technical performance metrics within RPZs. Utilizing a design science research methodology, the authors validated their framework through the creation of metrics that can be applied in real-world scenarios. This research is significant as it provides a structured approach to evaluating technical aspects of RPZ operations, which are crucial for maintaining security integrity and operational efficiency. The metrics discussed in this study include false-positive rate, mean time to detect, mean time to reaction, and threat actor attribution.

The 2015 study by Otsuka and colleagues (Otsuka et al., 2015) explored the human capital aspects of RPZs by identifying factors that contribute to analyst burnout and proposing a human capital model to mitigate this issue. Through anthropological study and grounded theory, the research involved fieldwork at both corporate and university RPZs. This study highlights the importance of addressing human factors to maintain high levels of performance and reduce turnover rates among RPZ analysts. The metrics discussed focus on analyst morale, operational efficiency, automation, and human capital management.

Enoch Agyepong and teams conducted a systematic review to identify the challenges faced by RPZ analysts and to review existing performance metrics. Their study covered literature published between 2008 and 2018, providing a comprehensive overview of

the evolving landscape of RPZ operations. The findings emphasized the need for robust performance metrics that can address the dynamic nature of security threats and operational demands. The metrics discussed include time to detect, time to respond, number of alerts analyzed, number of tickets closed, incident detection within a timeframe, and the quality of incident reports.

Finally, in 2025, Ichise and teams (Ichise et al., 2025) publish the new paper that proposed a systematic method for measuring the performance of RPZ analysts using the Analytic Hierarchy Process (AHP). This study involved a Delphi panel with industry experts and an experimental case study to validate the proposed approach. By focusing on a weighted approach to performance measurement, the research provides a nuanced and structured method for evaluating analyst performance, which is critical for continuous improvement in RPZ operations. The metrics discussed include the quality of incident analysis, quality of incident reports, number of incidents handled, time to resolve incidents, and the false positives rate. Table 2.4 explain the paper regarding related works with this works and their proposed methods :

Table 2.4 Comparison of Related Works in ISP Compliance and Negative Content Data Synchronization Metrics

No	Author	Objective	Methodology	Type of Metrics	Metrics Objectives	Key Metrics
1	(Sekar Asmoro Gati and Khoirunurrofik, 2023)	This research aims to investigate the impact of telecommunications infrastructure on inclusive economic growth in Indonesia. The study recognizes that different influences may arise in Indonesia compared to African countries, given its unique characteristics as a large archipelago nation and a leading emerging market in Asia.	Literature review, framework design, point-based metrics evaluation	Quantitative	Infrastructure, Operational, Technical	Telecommunications infrastructure on inclusive economic growth in Indonesia, utilizing panel data from 34 provinces spanning the years 2011 to 2021 and employing the Fixed Effect Model (FEM) methodology.
2	(Mallapragada et al., 2009)	The authors aim to design a metasemi-supervised learning algorithm that	Design science research methodology,	Quantitative and Qualitative	Technical	SemiBoost is a novel boosting framework for semi-supervised learning

No	Author	Objective	Methodology	Type of Metrics	Metrics Objectives	Key Metrics
		enhances the performance of any given supervised algorithm through the integration of unlabeled data. They propose a boosting framework, termed SemiBoost, which leverages pairwise similarity measurements to guide the selection of unlabeled examples and assign class labels during the learning process.	framework validation through metrics creation			designed to enhance the performance of various conventional supervised learning algorithms. It achieves this by leveraging a substantial volume of unlabeled data, demonstrating performance comparable to state-of-the-art semi-supervised learning approaches.
3	(Rahul et al., 2021)	Five machine learning models were applied, with Random Forest achieving the highest accuracy of approximately 87% in classifying various internet applications	Framework development based on analyst performance criteria	Quantitative and Qualitative	Operational, Technical	Five machine learning models were applied for traffic classification, with Random Forest achieving the highest accuracy of approximately 87%
4	(Durga et al., 2019)	The research surveys state-of-the-art machine and deep learning algorithms in the IoT healthcare domain.	Anthropological study, Grounded Theory, fieldwork at corporate and university RPZs	Quantitative and Qualitative	Management, Operational, Technical	Analyst morale, operational efficiency, automation, human capital management
5	(Utamima et al., 2023)	The study analyzes 100,000 reviews of the By.U application on Google Play Store using opinion mining techniques. It compares K-Nearest Neighbors, Support Vector Machine, and Naive Bayes algorithms for sentiment analysis	Systematic Literature Review (SLR)	Quantitative and Qualitative	Operational, Technical	The primary objective is to analyze public sentiment towards the By.U application on the Google Play Store. It compares the performance of K-Nearest Neighbors, Support Vector Machine, and Naive Bayes algorithms in sentiment analysis

No	Author	Objective	Methodology	Type of Metrics	Metrics Objectives	Key Metrics
		The By.U application frequently crashes, negatively impacting user experience and company reputation				
6	(Moreira and Fechine, 2018)	The paper presents a method to distinguish between bikini and pornographic images using machine learning techniques. It improves upon existing algorithms, achieving 96.96% accuracy and 94.31 F1-score	Design Science research methodology, framework validation through metrics creation	Quantitative and Qualitative	Management, Operational, Technical	The authors aim to improve existing detection algorithms for better accuracy in identifying inappropriate content. The research focuses on reducing false positives in portrait photos by incorporating detected faces as features
7	(Malathi and Ruby, 2024)	The study compares four machine learning models for predicting Deep Vein Thrombosis (DVT) risk in rehabilitation inpatients, Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Deep Neural Networks (DNN) are evaluated.	Random Forest (RF), Support Vector Machine (SVM), and Deep Neural Networks (DNN) are also employed for comparison	Quantitative and Qualitative	Management, Operational, Technical	Logistic regression is effective for DVT prediction, balancing simplicity and interpretability in clinical settings. The study emphasizes the importance of model explainability alongside performance metrics
8	(Nayak et al., 2023)	The paper discusses the integration of AI and ML in complex communication networks, particularly optical networks, to enhance performance and resilience. The research evaluates	The paper discusses various machine learning methods, including regression and classification, for data analysis in optical networks. It highlights the	Quantitative	The paper explores the integration of AI and ML in communication networks to enhance performance and address complexities. The research evaluates various ML methods and their effectiveness across different use cases. It	The paper concludes that integrating AI and ML is essential for enhancing communication networks' performance and addressing their complexities. The research discusses the applications of ML in network management and resilience, highlighting their effectiveness in ISP

No	Author	Objective	Methodology	Type of Metrics	Metrics Objectives	Key Metrics
		various ML methods and their effectiveness in network management and monitoring tasks	use of supervised and unsupervised learning to address regression and classification challenges		discusses the application of ML in optical networks for network management and resilience	
9	(Blanquero et al., 2021)	The paper discusses the integration of AI ML and paper proposes a sparse version of the Naïve Bayes classifier addressing feature correlation and classification performance constraints	The paper proposes a variable reduction method that samples certain combinations of features based on their dependencies, The proposed sparse Naïve Bayes classifier incorporates correlation structures and performance measures for feature selection	Quantitative	The paper aims to propose a sparse version of the Naïve Bayes classifier for datasets with correlated features. The objective is to demonstrate competitive results in accuracy, sparsity, and running times compared to existing methods	The Naïve Bayes classifier is applied in medicine, genetics, reliability, risk, and document analysis
10	(Bukhori et al., 2023)	The paper discusses sentiment analysis in learning evaluations using Artificial Intelligence and Naive Bayes classification methods. The	The research involves preprocessing and weighting stages for training data features classifier incorporates correlation structures and	Quantitative	The paper aims to propose a sparse version of the Naïve Bayes classifier for datasets with correlated features. The objective is to demonstrate competitive results in accuracy, sparsity, and running times	The Naïve Bayes classifier is applied in medicine, genetics, reliability, risk, and document analysis

No	Author	Objective	Methodology	Type of Metrics	Metrics Objectives	Key Metrics
		results indicate improved prediction accuracy in document form compared to methods without feature selection	performance measures for feature selection		compared to existing methods	
11	(Muslim, 2022)	The paper discusses Naive Bayes model is applied for HS Code classification to assist Customs Officials in tariff setting	The research focuses on using data mining techniques, specifically Naive Bayes classification, to improve HS Code classification accuracy	Quantitative	The study focuses on analyzing historical data to improve future decision-making. It aims to provide accurate classification predictions to mitigate potential state revenue loss	The study's findings have a high confidence level for estimating misclassification risk based on historical HS Code objections
12	(Paramartha et al., 2021) 30/08/2025 11:27:00	The paper discusses Random Forest and Multinomial NB classifiers are applied to improve classification accuracy of HS Codes	The study employs a machine learning framework to classify goods into HS Code based on DGCE's challenges	Quantitative	The paper discusses the imbalance in multi-class HS Code classification as a significant problem	Data preprocessing includes removing duplications, punctuations, stop words, non-English words, and lemmatization. Random sampling methods are used to select records from classes with varying sizes
13	(Ichise et al., 2023)	The paper discusses proposed system aims to analyze DNS traffic to identify direct outbound applications	The paper proposes a method to detect and block abnormal applications by analyzing DNS	Quantitative	The paper discusses system aims to detect and block various outbound applications based on DNS records	The discussion includes the limitations of existing DNS resolvers and their impact on application security RPZ

No	Author	Objective	Methodology	Type of Metrics	Metrics Objectives	Key Metrics
		without proper domain name resolutions.	traffic using DNS RPZ			
14	(Amalia et al., 2019)	The paper discusses addresses the massive spread of negative content on the internet, particularly in Indonesia, focusing on gambling, pornography, and fraud	The research employs supervised machine learning for website classification into categories like gambling and pornography	Quantitative	The research utilizes the TRUST+ database for classification based on keywords and website content	The research aims to classify negative content into categories like gambling, pornography, and fraud using existing resources from the TRUST+ database The Indonesian government issued Law No. 11 to regulate prohibited acts in cyberspace due to this problem.
15	(Mathrani and Alipour, n.d.)	The paper discusses methods for Internet filtering and ways to bypass these filters. The study investigates government censorship of websites across ten countries using the Website Block Finder application	The research employs keyword filtering to block website	Quantitative	The research investigates Internet filtering methods and bypass techniques at a national level.	The study investigates website filtering on the public Internet at the national level across ten countries. The research discusses commonly used methods for Internet filtering and bypassing techniques.

The reviewed literature underscores the critical role of metrics in evaluating and improving the performance of Response Policy Zones (RPZs). While existing research offers insights into various security metrics that can be applied to RPZ settings, there is still a noticeable gap in the literature regarding metrics specifically tailored for RPZs.

Although there are several innovative approaches proposed by RPZ researchers and vendors, such as real-time security level assessment, network resiliency simulation, and dynamic analyst scheduling models, the absence of standardized RPZ-specific metrics remains a significant challenge. These findings emphasize the urgent need for further academic investigation and development of metrics that are uniquely suited for RPZ environments specifically at XYZ. Addressing this gap will not only contribute to improving the effectiveness and efficiency of RPZ operations but also improve the overall cybersecurity posture. As such, future research efforts should focus on bridging this gap by systematically studying and establishing RPZ-specific metrics to better serve the needs of security practitioners and organizations.

In this thesis, we propose developing RPZ metrics that integrates technical and management perspectives. Traditionally, RPZ metrics have primarily focused on technical aspects, such as incident detection and response time, while neglecting broader management issues such as resource allocation and strategic alignment. Our model seeks to address this gap by describing metrics relevant to both technical and management objectives on compliance with negative content data.

2.21. Evaluation Methods

The studies reviewed employed various evaluation methods tailored to the challenges of classification, focusing on accuracy, F1-score, and cross-validation. Key findings include :

- (Paramartha et al., 2021): Random Forest and Multinomial Naive Bayes models were evaluated using F1-score, precision, and recall, with k-fold cross-validation ensuring reliability. While effective on clean, multilingual data, the models struggled with unclear descriptions.
- (Altaheri and Shaalan, 2020) : Accuracy was used to evaluate SVM and Naive Bayes models, which performed well on simple descriptions but struggled with hierarchical HS codes. Cross-validation was mentioned but not elaborated, limiting robustness for complex classifications.

- (Muslim, 2022) : The Naive Bayes model achieved 99.97 persen accuracy and a Kappa coefficient of 0.999 with structured data. However, it faced challenges with ambiguous and incomplete product descriptions, highlighting the limitations of real-world data.

To ensure the best possible performance of the model, it is important to choose the right evaluation metrics that align with the classification goals, particularly given the imbalanced datasets there are :

1. Accuracy is commonly used but may not be sufficient in cases of class imbalance.
2. F1 score is a crucial metric for balancing precision (the correctness of positive predictions) and recall (the ability of the model to find all relevant instances). In classification, this is particularly important when dealing with imbalanced data, as it avoids favoring the majority class while penalizing false positives and false negatives.
3. The precision-recall curve provides more detailed insights into how well the model performs across different thresholds, which is important when the dataset has a high-class imbalance.
4. A confusion matrix can be used to visualize the performance of the model in a more granular way, helping to identify misclassifications and providing insights into which specific categories or parts are being misclassified.

One of the issues of concern in machine learning models is interpretability. Ideally, every researcher expects not only optimal model performance but also ease of understanding. For several reasons, models are generally much more reliable if we understand the underlying factors behind their decision-making (Hidalgo-Fort et al., 2023).

Feature Importance is a measure that indicates the importance or contribution of a variable to changes in model performance. One measure that can be used is Permutation Importance. Permutation Importance measures the magnitude of the change in model performance when the values of a particular variable are randomly permuted (Altmann,

2010). Permutation Importance is model-agnostic, so it can be applied to various modeling algorithms.

SHAP (SHapley Additive exPlanations) Analysis is a game theory-based approach that explains the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values. For each prediction, SHAP assigns each feature an importance value representing its contribution to pushing the prediction away from the baseline (average) prediction.

The intuition behind permutation importance is that if the values of a variable are randomized but the model's performance remains unchanged, this indicates that the value of the variable has little impact. Conversely, if the model's performance experiences a significant decrease, it can be assumed that the value of the variable significantly impacts the model's predictive ability.

