information

MDPI

Article

A Hybrid Machine Learning Approach for Classifying Indonesian Cybercrime Discourse Using a Localized Threat Taxonomy

Firman Arifman *, Teddy Mantoro and Dini Oktarina Dwi Handayani

School of Computer Science, Nusa Putra University, Jakarta 43152, Indonesia;
teddy.mantoro@nusaputra.ac.id (T.M.); dini.oktarina@nusaputra.ac.id (D.O.D.H.)
* Correspondence: firman.arifman@nusaputra.ac.id

Abstract

Indonesia's rapid digital growth has been accompanied by escalating cyber threats, with public discourse on social media emerging as a critical but underutilized source of threat intelligence. This discourse is characterized by informal language and local nuances that render existing international cybercrime taxonomies ineffective, creating a gap in scalable, locally relevant threat analytics. This study introduces the Indonesian Cybercrime Threat Taxonomy (ICTT), a novel five-dimensional framework tailored to Indonesian online environments. An end-to-end OSINT pipeline was developed to collect 2344 samples from X (formerly Twitter) and YouTube, employing weak supervision with 12 high-precision regex patterns to generate training labels. A state-of-the-art IndoBERT model was fine-tuned on this data, and its performance was compared against rule-based and hybrid classification models. On a manually annotated gold-standard dataset of 600 samples, both the IndoBERT and hybrid models achieved 96.8% accuracy, significantly outperforming the rule-based baseline (66.7%). The models demonstrated strong generalization across both social media platforms, and the hybrid approach provided an effective balance of high performance and interpretability. This research demonstrates that informal public discourse can be systematically transformed into structured threat intelligence. The ICTT and the accompanying hybrid classification system provide a scalable, interpretable, and locally relevant foundation for cyber threat analytics in Indonesia, establishing a methodological blueprint for other low-resource language contexts.

Keywords: cybercrime classification; machine learning; taxonomy; natural language processing; weak supervision; transformer models

Check for updates

Academic Editors: Jennifer D'Souza and Chengzhi Zhang

Received: 12 February 2026
Revised: 8 March 2026
Accepted: 18 March 2026
Published: 20 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The Indonesian digital economy has experienced explosive growth over the past decade, with internet penetration reaching approximately 80% of the population in 2025 [1]. According to Bank Indonesia (BI), total digital payments in 2024 reached approximately 34.5 billion transactions across digital channels. This includes payments via mobile, internet banking, Quick Response Code Indonesian Standard (QRIS), and other digital payment systems. However, this rapid digitalization has been accompanied by a proportional surge in malicious cyber activities. According to the Indonesian Cybersecurity Agency (BSSN), Indonesia recorded 3.64 billion cyberattacks in the first half of 2025 alone, representing a significant escalation in threat volume and sophistication [2,3]. The financial impact is

substantial. According to Indonesia's Financial Services Authority (Otoritas Jasa Keuangan, OJK), public reports handled through the Indonesia Anti-Scam Centre (IASC) between its launch on 22 November and the end of 2024 recorded total financial losses of Rp309.5 billion (approximately USD 19.6 million), of which only the loss of Rp83.6 billion (around USD 5.3 million; 27.01%) was successfully prevented, highlighting the significant economic harm and limited recovery associated with cyber-enabled fraud in Indonesia [4].

A critical yet largely untapped source of threat intelligence exists in the vast, unstructured public discourse on social media platforms such as X (formerly Twitter) and YouTube. Indonesian users frequently share real-time accounts of cybercrime incidents, phishing attempts, scams, and malware infections using informal language, local slang, and colloquial expressions. This abundance of user-generated content offers a rich, authentic perspective on actual threat landscape as it is experienced by the general population. However, the potential of this discourse remains unrealized due to a fundamental challenge: the language is informal, contextually specific to Indonesia, and lacks the structured terminology found in formal incident reports [3].

Existing international cybercrime taxonomies, such as those developed by the European Union Agency for Cybersecurity (ENISA) [5], the Internet Organised Crime Threat Assessment (IOCTA) [6], the Verizon Data Breach Investigations Report (DBIR) [7], and the MITRE ATT&CK framework [8], provide widely adopted structures for categorizing threats. However, these frameworks were designed primarily for formal, structured incident reports from institutional sources and do not align well with the unique linguistic and regulatory characteristics of the Indonesian online environment. This creates a critical gap: there is no standardized, validated method to capture, structure, and analyze the informal cybercrime narratives populating Indonesia's digital public sphere.

Problem definition: Let D be a corpus of informal Indonesian social media posts relating to cybercrime. The primary objective of this research is to learn a mapping function f: D $\rightarrow$ Y, where Y is the set of threat labels defined by the ICTT. The successful development of such a function is contingent upon addressing the following four key challenges, which is the core focus of this analytical study.

1. High-Dimensional and Sparse Vocabulary: The language used on Indonesian social media is informal, rich in slang, and contains many out-of-vocabulary terms not found in standard NLP corpora.
2. Significant Label Noise: The inherent ambiguity of informal language makes it difficult to assign clear, unambiguous labels, even for human annotators.
3. Severe Class Imbalance: Reports of some threat types (e.g., WhatsApp Phishing) are far more prevalent in the public discourse than others (e.g., Ransomware), leading to a highly imbalanced dataset.
4. Interpretability vs. Performance Trade-Off: To be useful for threat analysts and policymakers, the system must not only be accurate but also interpretable.

In addressing these challenges, this study offers three core contributions:

1. Design of the Indonesian Cybercrime Threat Taxonomy (ICTT): The ICTT is a novel, five-dimensional framework tailored to the nuances of Indonesian cybercrime discourse, bridging formal policy language with informal citizen language.
2. An End-to-End OSINT and Machine Learning Pipeline: This study offers a complete system for collecting, preprocessing, weakly labeling, and classifying Indonesian cybercrime content, incorporating state-of-the-art NLP techniques.
3. Comparative Analysis of Classification Approaches: This study conducted a rigorous evaluation of rule-based, transformer-only, and hybrid classification models, demonstrating the effectiveness of combining interpretability with deep learning performance.

This paper is organized as follows: Section 2 describes a review of related studies on cybercrime taxonomies, NLP-based threat detection, weak supervision, and hybrid classification systems. Section 3 details the methodology from the development of the ICTT to the architecture and training of the classification models. Section 4 presents the comprehensive results of the comparative evaluation, including per-class analysis and platform-specific performance. Section 5 concludes the paper.

2. Related Work

Other works reviewed in this study cover the intersection of cybercrime classification, natural language processing, weak supervision, and hybrid machine learning systems.

2.1. Cybercrime Taxonomies and Threat Classification

Standardized taxonomies are fundamental for the systematic analysis of cyber threat landscapes, as they facilitate consistent communication among cybersecurity practitioners, policymakers, and researchers. The ENISA Threat Taxonomy [5] offers a comprehensive framework for structuring threat information by categorizing incidents according to attack type, target, and impact. Complementing this, the Verizon Data Breach Investigations Report (DBIR) [7] provides empirical insights into real-world breach patterns derived from thousands of documented incidents. Furthermore, the MITRE ATT&CK framework [8] serves as an extensive knowledge base of adversary tactics and techniques based on practical observations.

Despite their utility, these frameworks exhibit significant limitations when applied to informal, user-generated cybercrime discourse. Chandra and Snowe [9] highlight the need for localized taxonomies that reflect specific regional contexts, regulatory environments, and linguistic nuances. Additionally, research by Agrafiotis et al. [10] on cyber-harm taxonomies emphasizes the importance of understanding how impacts propagate through diverse social and organizational settings. Analytical work by Malavasi et al. [11] further demonstrates that existing taxonomies often possess substantial structural and semantic variations, complicating their universal application.

The Indonesian Cybercrime Threat Taxonomy (ICTT) addresses these gaps by incorporating threat types, attack vectors, and victim categories specific to the Indonesian landscape. While maintaining alignment with international standards where appropriate, the ICTT integrates localized regulatory contexts, such as the Electronic Information and Transactions Law (UU ITE).

Notwithstanding the breadth of extant research, no prior study has developed a classification framework specifically tailored to the Indonesian cybercrime environment. International frameworks, including ENISA [5], IOCTA [6], DBIR [7], and MITRE ATT&CK [8], were designed primarily for Western or global threat landscapes. Consequently, they often fail to account for distinctive threat patterns prevalent in Indonesia, such as WhatsApp-based phishing, SIM swap attacks targeting domestic mobile networks, and e-wallet fraud exploiting local platforms like OVO, GoPay, DANA, QRIS, and LinkAja.

Academic taxonomies proposed by Chandra and Snowe [9] and Agrafiotis et al. [10] similarly focus on generalized cybercrime categories without localizing to specific national or linguistic contexts. To the best of the authors' knowledge, the ICTT introduced in this study is the first taxonomy specifically designed to address the Indonesian threat landscape. It draws upon Indonesian regulatory frameworks, local platform ecosystems, and a qualitative analysis of Indonesian social media discourse to provide a more granular and relevant classification system.

2.2. Natural Language Processing for Cybersecurity

The application of NLP techniques to cybersecurity has increased substantially in recent years. Arazzi et al. [12] provide a comprehensive overview of NLP-based techniques for cyber threat intelligence, including text classification, entity recognition, and relationship extraction. Albarrak et al. [13] discuss natural language processing frameworks for analyzing cyber threat intelligence from unstructured data, such as vulnerability reports, social media chats, security advisories, and technical forums. Building on this foundation, Transformer-based models such as BERT, RoBERTa, and XLNet [14] significantly improved classification performance by modeling full contextual relationships.

Domain-specific language models have proven particularly effective for cybersecurity applications. Bayer et al. [15] introduced CySecBERT, a domain-adapted language model for the cybersecurity domain, demonstrating that pre-training on domain-specific corpora significantly improves performance on cybersecurity-related NLP tasks. Aghaei et al. [16] introduced SecureBERT, a domain-specific language model tailored for cybersecurity designed to capture nuanced semantic patterns in cybersecurity-related texts (e.g., cyber threat intelligence). The model enables automation of various critical cybersecurity tasks that would otherwise depend heavily on human expertise and time-intensive manual analysis.

For Indonesian-language NLP, Koto et al. [17] introduced IndoBERT, a pre-trained bidirectional transformer model specifically designed for Indonesian language understanding. IndoBERT was trained on a large corpus of Indonesian text from diverse sources and has demonstrated state-of-the-art performance on multiple Indonesian NLP benchmarks [18]. The availability of IndoBERT makes it feasible to develop sophisticated NLP systems for Indonesian cybercrime discourse.

2.3. Weak Supervision and Noisy Label Learning

The bottleneck of creating large-scale, high-quality labeled datasets has motivated the development of weak supervision techniques. Ratner et al. [19] formalized the weak supervision paradigm through the Snorkel framework, which enables rapid training data creation using multiple noisy labeling functions. Their key insight was that while individual labeling functions may be imprecise, their combination through principled aggregation can produce training labels of sufficient quality for downstream machine learning models.

Recent work has demonstrated that modern deep learning models, particularly transformer-based models like BERT, are surprisingly robust to label noise. Zhu et al. [20] conducted a comprehensive study on BERT's robustness to label noise in text classification, finding that BERT can learn effectively even when trained on datasets with significant label noise, often outperforming more complex noise-handling techniques. This finding is particularly relevant for this work, as it suggests that the noisy weak labels generated by the rule-based system can serve as effective training signals for IndoBERT, helping to mitigate the need for extensive manual data annotation.

2.4. Hybrid Rule-Based and Machine Learning Systems

Combining rule-based systems with machine learning models is an established strategy for text classification and other NLP tasks. Villena-Román et al. [21] demonstrated that hybrid approaches combining machine learning algorithms with rule-based expert systems can achieve superior performance compared to either approach alone, particularly when interpretability and transparency are important requirements.

Li et al. [22] developed a hybrid medical text classification framework that integrates attentive rule construction with neural networks, showing that explicit rule-based components can enhance model interpretability while maintaining high performance. The key advantage of hybrid systems is that they provide a balance between the transparency and

interpretability of rule-based approaches and the generalization power and performance of machine learning models. In the cybersecurity domain, hybrid approaches are particularly valuable because they enable human analysts to understand and audit the decision-making process, which is critical for operational deployment and regulatory compliance. The application of these advanced NLP and hybrid techniques is crucial for understanding emerging threats and developing robust defense mechanisms, from exploring hacker assets for proactive intelligence [23] and surveying the use of deep learning [24] to characterizing specific threats like Ransomware as a Service (RaaS) [25] and building scalable cyber threat analysis systems [26].

3. Methodology

The methodological design of this study relies on multiple layers that interact with one another rather than forming a single uninterrupted pipeline. At its core, the approach combines structured taxonomy work with empirical modeling of cybercrime discourse drawn from public platforms. The intention is not to claim that one analytic technique is inherently superior but to recognize that different parts of the problem may benefit from different tools.

A recurring issue when dealing with Indonesian cybercrime narratives is that they do not present the same linguistic shape across platforms. Some platforms reward quick, compact expression, while others create space for more elaborate descriptions. These variations may influence how threat indicators surface in text and how reliably they can be detected. For this reason, the ICTT is used as the conceptual anchor of the entire system. It helps maintain internal consistency when classifying text from very uneven sources and guides later evaluation by making it easier to determine where errors cluster.

The overall research workflow, illustrated in Figure 1, comprises six primary phases adapted from the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework [27]. Despite the emergence of specialized methodologies for Big Data, CRISP-DM remains the most widely adopted process model in both academic research and industrial practice due to its robust, iterative nature [28,29]. The adapted phases are (1) ICTT Construction, (2) OSINT Data Acquisition, (3) Preprocessing and Gold Labeling, (4) Modeling, (5) Evaluation and Comparative Analysis, and (6) Deployment and Dashboard.

Although these steps are presented sequentially, the operational workflow remains highly flexible. For instance, errors observed during evaluation frequently necessitate revisions in the preprocessing stage, while the annotation process may reveal conceptual gaps in the taxonomy that necessitate further refinement. Furthermore, inconsistencies identified during dashboard deployment often require a return to earlier modeling or data preparation stages. Consequently, the research design evolves through continuous refinement rather than strictly linear progression.

This iterative arrangement acknowledges that cybercrime discourse rarely conforms to rigid, predefined templates. This research approach oscillates between definition, testing, and revision to ensure empirical accuracy. The hybrid modeling architecture developed is a direct outcome of this pattern; it emerged not as a predetermined selection but as a practical response to the divergent performance of rules and machine learning models across various categories of Indonesian cybercrime text.

3.1. The Indonesian Cybercrime Threat Taxonomy (ICTT)

The ICTT is a five-dimensional framework designed to provide a holistic view of cybercrime incidents as described in informal text, enabling systematic interpretation and structured representation of heterogeneous narrative data. Its design was guided by (1) an analysis of existing international taxonomies (ENISA, IOCTA, DBIR, and MITRE

ATT&CK), (2) a review of Indonesian cybersecurity policy and regulatory frameworks (UU ITE-Law No. 11 of 2008 on Information and Electronic Transactions), (3) a qualitative analysis of 600 Indonesian social media posts describing cybercrime incidents, and (4) consultations with cybersecurity domain experts to ensure contextual relevance and practical applicability.

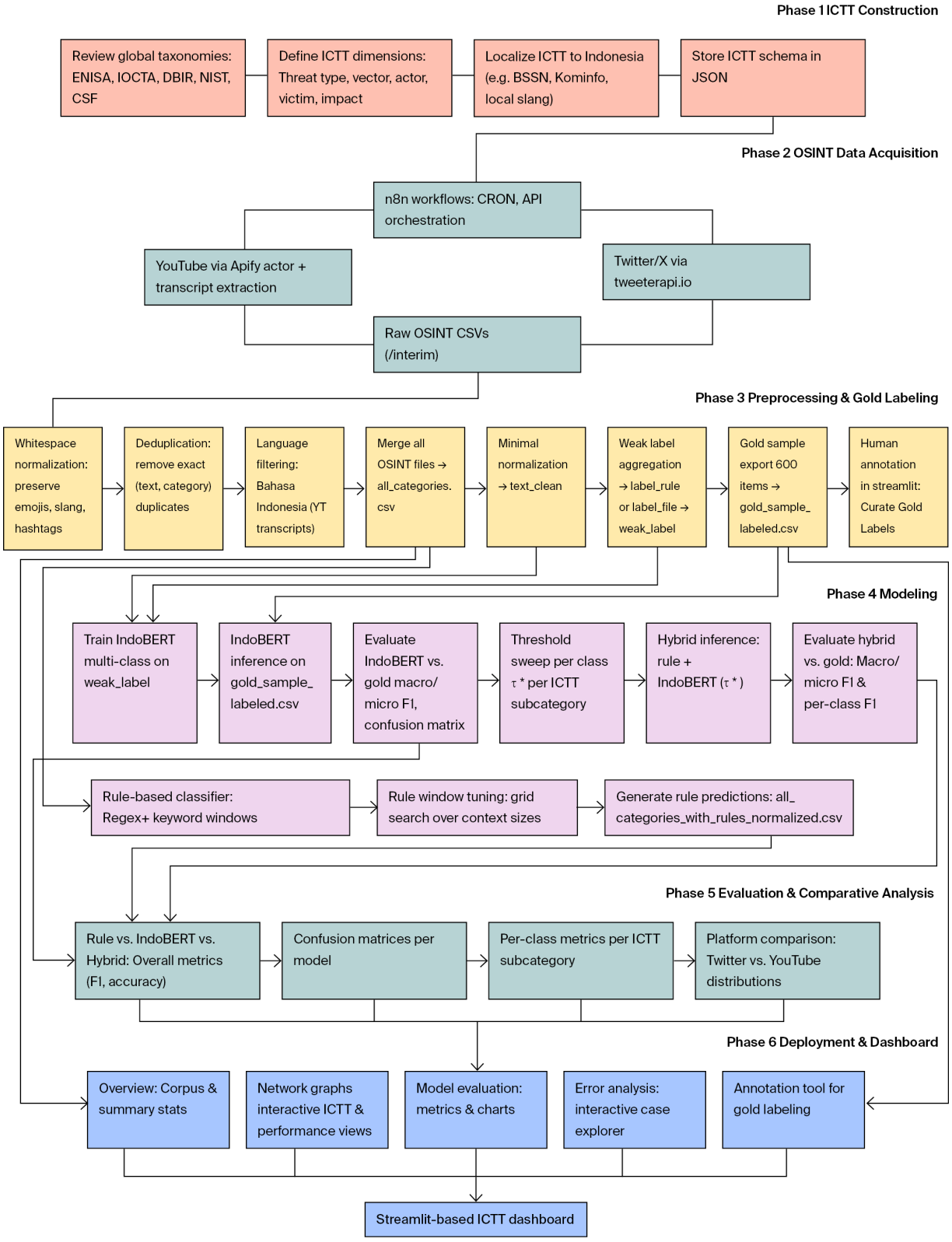


Figure 1. End-to-end research methodology.

The five dimensions of the ICTT were grounded in an analysis of Indonesian cybercrime social media discourse and designed to evolve as new threat patterns emerge. These dimensions were systematically derived through thematic coding of 600 social media posts, where recurring linguistic patterns were mapped to conceptual categories. This data-driven approach revealed that Indonesian cybercrime discourse consistently references the type of attack, the delivery method, the perpetrator, the affected parties, and the resulting harm. These dimensions were subsequently validated against international standards and refined through expert consultation. Each dimension includes an "Other" category for types not yet represented.

1. Threat Type—the nature and category of the malicious act. The ICTT enumerates 10 major threat categories with over 60 specific subcategories (e.g., Phishing & Social Engineering, Malware & Malicious Software, Fraud & Online Scams, Data Breach & Identity Theft, Hacking & System Intrusion, Financial & Payment Attacks, Online Child Exploitation, Cyber Harassment & Online Abuse, Cyber-Enabled Traditional Crime, and Emerging Threats). New categories will be added when 50 or more validated samples describe a distinct, operationally relevant threat type.
2. Attack Vector—the delivery mechanism or channel through which the threat is executed. The ICTT identifies five primary attack vectors (Messaging/Apps, such as WhatsApp and Telegram; Social Platforms, such as X, Instagram, Facebook, TikTok, and YouTube; Email/SMS/Voice; Web & Apps; and Network/Infrastructure). The prominence of messaging and social platform vectors reflects their centrality in Indonesian online communication and cybercrime reporting.
3. Threat Actor—the entity or group responsible for perpetrating the cybercrime. The ICTT classifies threat actors into three categories (non-state actors, including cybercriminal groups and hacktivists; Internal/Partner actors, such as insiders and contractors; and State-Linked actors, including Advanced Persistent Threats and state-sponsored operations).
4. Victim—the target or affected entity. The ICTT enumerates four prevalent victim categories (Private Sector, including SMEs and e-commerce platforms; Finance, including banks and fintech companies; Public Sector, including government agencies and educational institutions; and Individuals, including citizens and minors). While the Indonesian Cybersecurity Agency (BSSN) 2024 Cybersecurity Landscape report [3] identifies additional victim categories in critical infrastructure incidents (e.g., IoT/CCTV systems, healthcare institutions), references to them are not prevalent in current social media discourse and may emerge in future analyses.
5. Impact—the consequences or harm resulting from the incident. The ICTT classifies incident impact into four dimensions (Confidentiality impacts, such as Privacy/Data Protection; Integrity/Availability impacts, such as Service Disruption/Outage; Financial/Legal impacts, such as Financial Loss and Regulatory Exposure; and Reputation/Societal impacts, such as Reputation Damage and Public Safety/National Security concerns).

This multi-dimensional structure, illustrated in Figure 2, allows for a more nuanced classification than single-label approaches, capturing the complexity of real-world cybercrime incidents.

Table 1 reveals that while the ICTT shares conceptual overlap with ENISA and MITRE ATT&CK for standard threat types (ransomware, email phishing, and credential theft), 8 of the 32 ICTT subcategories (25.0%) are Indonesia-specific and lack direct equivalents in either international framework. These include WhatsApp Phishing and SMS/Phone phishing, which exploit the dominant communication channels in Indonesia; SIM Swap attacks, which target the Indonesian mobile ecosystem; E-wallet fraud targeting Indonesian fintech

platforms (OVO, GoPay, DANA, and LinkAja); Loan Scams exploiting Indonesia's informal lending market; Online Gambling and Narcotics Trade, reflecting cyber-enabled versions of traditional crimes prevalent in Indonesia; and Deepfake Scams, an emerging threat type not yet formally addressed in ENISA or MITRE ATT&CK. These eight subcategories collectively constitute the Indonesia-specific contribution of the ICTT and justify its development as a localized taxonomy tailored to the Indonesian threat landscape rather than a generic adaptation of international frameworks.

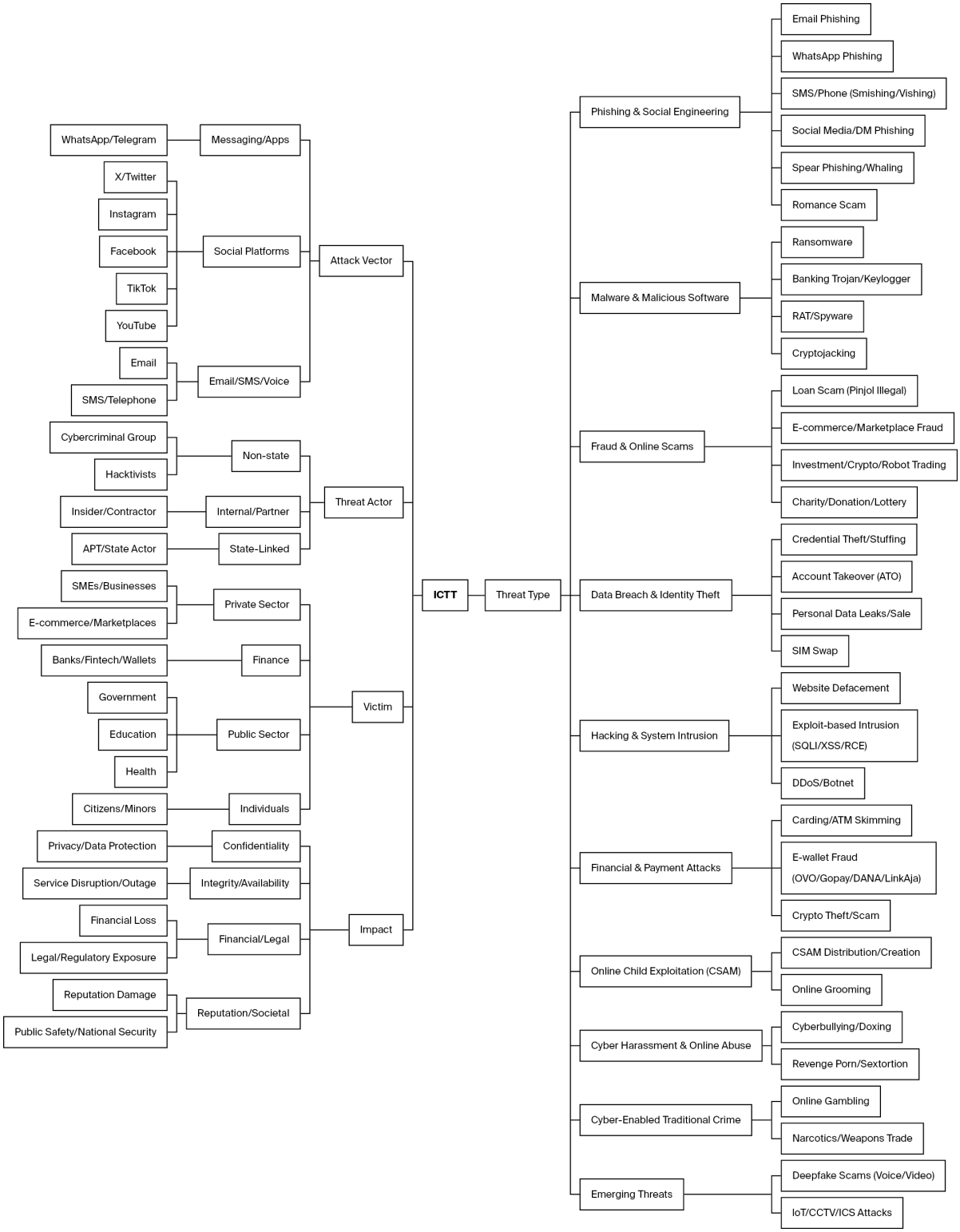


Figure 2. The Five dimensions of the ICT Taxonomy.

Table 1. Structural comparison of ICTT with ENISA and MITRE ATT&CK taxonomies.

ICTT Dimension	ICTT Subcategory	ENISA Equivalent	MITRE ATT&CK Equivalent	Indonesian-Specific?
Phishing & Social Engineering	Email Phishing	Phishing/Social Engineering	T1566 Phishing	No
	WhatsApp Phishing	Phishing (platform-specific)	T1566 (partial)	Yes
	SMS/Phone (Smishing/Vishing)	Phishing/Social Engineering	T1566 (partial)	Yes
	Social Media/DM Phishing	Phishing/Social Engineering	T1566 (partial)	No
	Spear Phishing/Whaling	Phishing/Social Engineering	T1566 Phishing	No
	Romance Scam	Social Engineering/Fraud	T1566 (partial)	No
Malware & Malicious Software	Ransomware	Malware/ Ransomware	T1486 Data Encrypted for Impact	No
	Banking Trojan/Keylogger	Malware	T1056 Input Capture	No
	RAT/Spyware	Malware	T1005 Data from Local System	No
	Cryptojacking	Malware	T1496 Resource Hijacking	No
Fraud & Online Scams	Loan Scam (Illegal Online Loan)	Fraud	No direct mapping	Yes
	E-commerce/Marketplace Fraud	Fraud	No direct mapping	No
	Investment/Crypto/Robot Trading	Fraud	No direct mapping	No
	Charity/Donation/Lottery	Fraud	No direct mapping	No
Data Breach & Identity Theft	Credential Theft/Stuffing	Identity Theft/Credential Compromise	T1110 Brute Force	No
	Account Takeover (ATO)	Identity Theft	T1110 Brute Force	No
	Personal Data Leaks/Sale	Information Disclosure	T1041 Exfiltration	No
	SIM Swap	Identity Theft (emerging)	No direct mapping	Yes
Hacking & System Intrusion	Website Defacement	Web-Based Attacks	T1491 Defacement	No
	Exploit-Based Intrusion (SQLi/XSS/RCE)	Web-Based Attacks	T1190 Exploit Public-Facing Application	No
	DDoS/Botnet	Denial of Service	T1498 Network Denial of Service	No
Financial & Payment Attacks	Carding/ATM Skimming	Fraud/ Payment Fraud	No direct mapping	No
	E-wallet Fraud (OVO/Gopay/DANA/LinkAja/QRIS)	Fraud (platform-specific)	No direct mapping	Yes
	Crypto Theft/Scam	Fraud	No direct mapping	No
Online Child Exploitation (CSAM)	CSAM Distribution/Creation	Information Disclosure/Abuse	No direct mapping	No
	Online Grooming	Social Engineering/Abuse	No direct mapping	No

Table 1. *Cont.*

ICTT Dimension	ICTT Subcategory	ENISA Equivalent	MITRE ATT&CK Equivalent	Indonesian-Specific?
Cyber Harassment & Online Abuse	Cyberbullying/Doxing	Information Disclosure/Abuse	No direct mapping	No
	Revenge Porn/Sextortion	Information Disclosure/Extortion	No direct mapping	No
Cyber-Enabled Traditional Crime	Online Gambling	Fraud	No direct mapping	Yes
	Narcotics/Weapons Trade	Fraud/Illegal Activity	No direct mapping	Yes
Emerging Threats	Deepfake Scams (Voice/Video)	Social Engineering (indirect)	No direct mapping	Yes
	IoT/CCTV/ICS Attacks	System Failures/Malware	T1200 Hardware Additions	No

Note: "Partial" indicates that the international taxonomy covers the general threat class but lacks specificity for the Indonesian context, platform, or regulatory dimension. "No direct mapping" indicates that the threat type is absent from the international taxonomy. "Emerging" indicates that the threat type is recognized but not yet formally integrated into the taxonomy.

3.2. Data Collection and Preprocessing

An OSINT pipeline was developed to collect a dataset of 2344 samples from X (n = 1875) and YouTube (n = 469) between 19 October 2025 and 10 November 2025, using their respective APIs. The selection of X and YouTube as the primary data sources was a deliberate methodological decision based on three criteria: (1) both platforms offer publicly accessible APIs, enabling systematic and reproducible OSINT collection; (2) they are among the most widely used platforms for cybercrime-related discourse in Indonesia, with X being particularly prevalent for real-time incident reporting and YouTube for victim testimonials and awareness content; and (3) they represent complementary linguistic registers, consisting of short, informal posts on X versus longer, more structured YouTube content. This variety provides linguistic diversity that strengthens the generalizability of the classification models.

Platforms such as Instagram, TikTok, and Facebook were considered but excluded due to API access limitations that would compromise the reproducibility and scalability of the collection pipeline. Data collection was conducted over a three-week window to ensure temporal coherence, capturing a consistent snapshot of the threat landscape without confounding effects from major policy changes or platform algorithm updates [30]. Keywords for data collection were derived from the ICTT, including terms such as "akun diretas" (account hacked), "kena tipu" (got scammed), "saldo ilang" (balance disappeared), and "phishing." A total of 2344 unique samples were collected. To enable training of supervised classification models without manual annotation of the entire dataset, a two-stage weak supervision strategy was employed to assign labels to the 2344 mined samples:

1. File-based labeling—Samples were retrieved from X and YouTube using threat-specific keyword queries. For instance, WhatsApp Phishing samples were collected using Indonesian terms such as "wa kena hack" (WhatsApp hacked), "link WA palsu" (fake WhatsApp link), and" hadiah wa" (WhatsApp prize). Labels were assigned based on the keyword set used for retrieval. While this ensures broad coverage, it assumes that keyword-based mining accurately reflects the threat type, a premise often complicated by semantic ambiguity in the Indonesian language.
2. Rule-based labeling—To improve label precision, a set of 12 core regular expression (regex) patterns was developed for the four most prevalent threat categories (WhatsApp Phishing, Email Phishing, Deepfake Scams, and Ransomware). These

patterns encode linguistic markers specific to Indonesian cybercrime discourse and were applied to all 2344 mined samples. Samples matching one or more rules received a rule-based label; samples with no matches retained only their file-based label or remained unlabeled.

The preprocessing pipeline consisted of four primary stages: (1) merging disparate files into a single unified table and reshaping columns into a consistent schema for downstream processing; (2) deduplicating posts to ensure data integrity; (3) performing text normalization, including whitespace trimming, spacing consolidation, and line-break standardization; and (4) assigning weak labels to each textual instance.

3.3. Data Characteristics and Representational Biases

The 2344 weakly labeled samples and 600 gold-standard samples exhibit several representational biases that warrant explicit documentation, as they have direct implications for model performance and generalizability. Understanding these biases is essential for interpreting results and assessing the fairness of the classification system.

3.3.1. Class Distribution and Imbalance

The gold-standard dataset exhibits significant class imbalance among threat categories. Among all samples, 46 were instances of WhatsApp Phishing (7.7% of threat-positive samples), 39 were Deepfake Scams (6.5%), 30 were instances of Email Phishing (5.0%), and only 9 represented Ransomware (1.5%). This distribution reflects the actual prevalence of threat types in Indonesian social media discourse, where WhatsApp-based phishing attacks are substantially more prevalent than ransomware incidents. However, the imbalance creates a 5.1:1 ratio between the most frequent (WhatsApp Phishing) and least frequent (Ransomware) threat categories. To mitigate the effects of this imbalance on model training, inverse-frequency class weighting, where each class weight is inversely proportional to its frequency, was applied when fine-tuning IndoBERT. This weighting strategy ensured that the model does not learn to favor high-frequency classes, but the imbalance remains evident in per-class performance metrics (see Section 4.3).

3.3.2. Platform Representation Bias

The dataset exhibits platform representation bias, with 75.7% of samples (454 samples) originating from X and 24.3% (146 samples) from YouTube. This imbalance reflects the higher volume of real-time cybercrime discourse on X but introduces a risk of platform-specific overfitting. The linguistic characteristics of the two platforms differ substantially: X content is typically short (280 characters), informal, and slang-heavy, while YouTube comments and video descriptions are longer, more structured, and often more explanatory. To assess whether the model generalizes across these linguistic differences, platform-specific evaluation was conducted separately for X and YouTube (see Section 4.4), revealing that IndoBERT achieves nearly identical macro-F1 scores on both platforms (0.973 on X; 0.803 on YouTube) despite the platform distribution imbalance. However, the YouTube evaluation subset is substantially smaller (16 threat-positive samples vs. 109 on X), which limits the statistical confidence in cross-platform generalization claims and necessitates cautious interpretation of YouTube-specific results.

3.3.3. User Representation Bias

The dataset exhibits user representation bias in two dimensions. First, the distribution of posts across users is highly skewed: while 443 unique users contributed to the 600-sample gold-standard dataset (1.35 posts per user on average), the top user (‘grok’) contributed 36 samples (6.0% of the entire dataset), and the top 15 users contributed 19.5%

of the data. The user 'grok' appears to be an automated or semi-automated account posting cybersecurity-related content, and its disproportionate representation may introduce systematic biases in the model's learned patterns, potentially overweighting the communication style and threat perceptions of this single source. Second, the dataset is heavily skewed toward personal/individual accounts (572 samples, 95.3%) compared to official or institutional accounts such as police departments, cybersecurity agencies, and financial regulators (28 samples, 4.7%). This skew means that the model is optimized for citizen-reported threats rather than institutional threat reporting, potentially limiting its applicability to formal incident response workflows or integration with official cybersecurity agencies.

3.3.4. Implications for Model Fairness and Deployment

These representational biases have direct implications for model fairness and real-world deployment. The class imbalance means that the model exhibits higher sensitivity to detecting WhatsApp Phishing (F1 = 0.968) compared to Ransomware (F1 = 0.800), which may result in alert fatigue for common threats while potentially missing rare but critical attacks. The platform bias means the model's performance on emerging platforms (e.g., TikTok, Instagram) is uncertain and requires additional evaluation. The user representation bias means that the model reflects the threat perceptions and communication patterns of individual Indonesian users rather than institutional cybersecurity professionals, which may limit its utility for formal threat intelligence workflows. These limitations are acknowledged and addressed through the evaluation methodology (platform-specific assessment; per-class performance reporting) and inform the direction of future work (see Section 5).

3.4. Classification Framework

This study designed and evaluated three classification models, as described below.

3.4.1. Rule-Based Classifier

The 12-core regular expression (regex) pattern rule-based classifier is intentionally compact, with rules that are distributed across four high-signal threat categories, as mentioned previously.

The rule-based classification step enriches each record with additional fields that summarize for which rules hits were recorded and how strongly they support a particular ICTT subcategory. For each text, the script loads the rule configuration from `ictt_rules.json`, compiles the relevant regex patterns with the selected window sizes, and scans the text label by label. Matched rules are collected per subcategory and used to compute a simple confidence score based on the number of rule hits. If one or more labels register a match, the label with the highest confidence is assigned to `label_rule`, and the corresponding rule identifiers and confidence value are stored in `rule_ids` and `rule_confidence`. Records with no matches remain unlabeled by the rule system and are assigned a confidence of zero. The overall process is illustrated in Figure 3.

The structure of rules that are defined inside a JSON configuration file (`ictt_rules.json`) keeps the rules interpretable. The configuration stores each subcategory as a separate group with a list of pattern definitions. A simplified excerpt illustrates how the WhatsApp phishing rules are structured:

```
    {
    ''id'':  ''wa_otp_link'', ''pat
tern'':''(?i)(wa|whats\\s*app|whatsapp).{0,10}(kode|otp|one[-
\\s]*time|verifikasi|login|masuk|
    akun)''
    }
```

This rule looks for short-range associations between WhatsApp references and common verification or login terms. The window of ten characters, indicated by {0,10}, limits the scope so that unrelated phrases are not clustered together. Although this is a small constraint, it prevents many false positives.

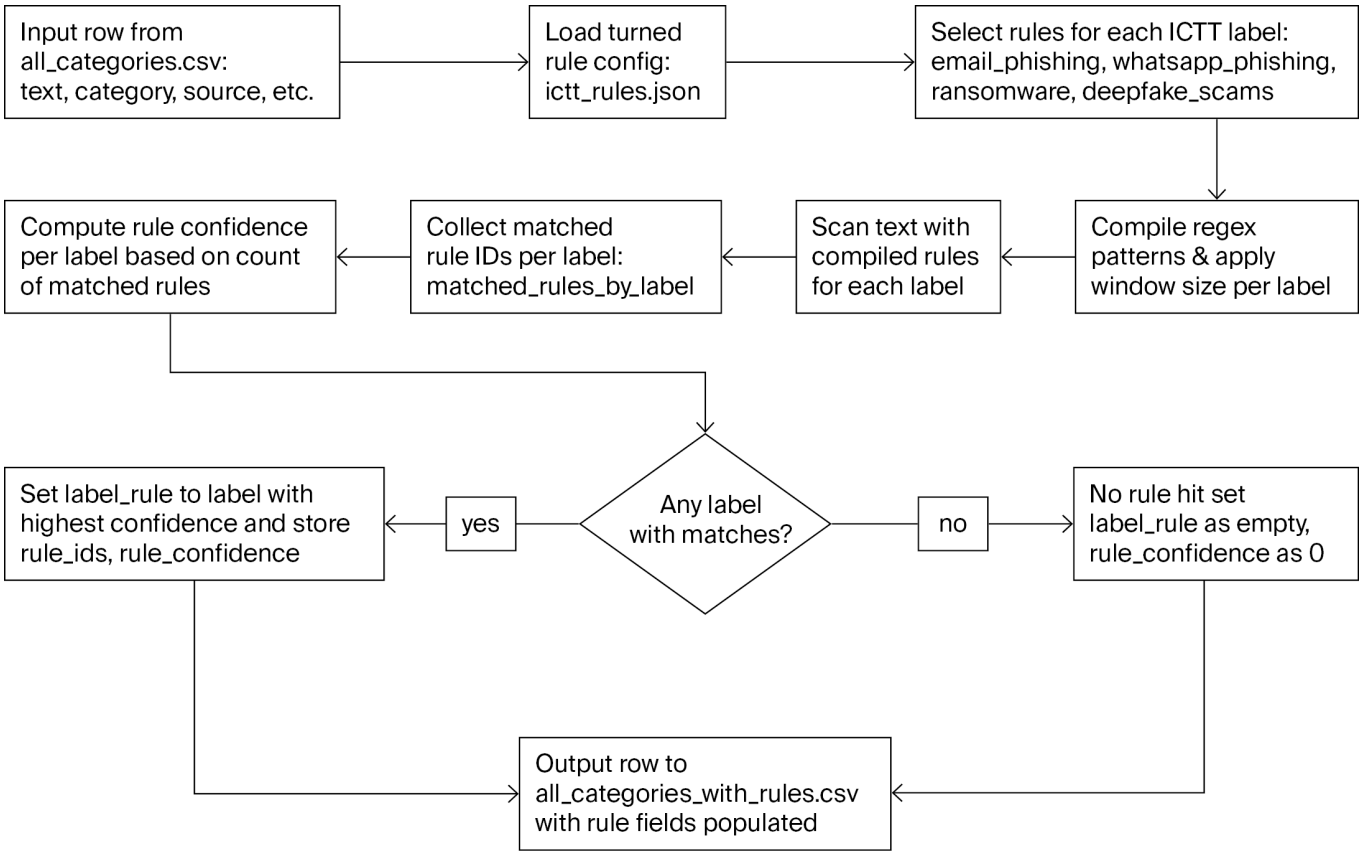


Figure 3. Rule-based side of the weak-labeling pipeline.

Window size plays a noticeable role in how well the rule-based classifier detects threat patterns. To quantify this effect, a sweep of candidate window values was evaluated for each ICTT subcategory using the gold-labeled dataset. The resulting tuning curve and data, presented in Figure 4 and Table 2, shows that performance does not improve uniformly as the window expands.

Table 2. Rule window tuning curve data.

ICTT Subcategory	Optimal Window Size (Characters)	Precision	Recall	F1-Score
deepfake_scams	20	0.471	0.800	0.593
email_phishing	10	0.352	0.833	0.495
ransomware	40	0.118	0.889	0.208
whatsapp_phishing	10	0.425	0.978	0.592

To characterize the noise in the weakly labeled training data and assess the reliability of both labeling sources, the 2344 mined samples were matched against the 600 gold-standard samples using a composite unique identifier (text content, username, and URL). This matching procedure yielded 781 samples that overlap with the gold set: 459 rule-labeled samples and 322 file-labeled samples. These matched samples enable direct validation of both weak supervision sources, as shown in Table 3.

Table 3. Weak supervision precision by source and threat category (validated against gold standard).

ICTT Subcategory	Rule-Labeled Samples (n)	Rule Precision	File-Labeled Samples (n)	File Precision
deepfake_scams	66	45.5% (30/66)	81	8.6% (7/81)
email_phishing	177	44.6% (79/177)	110	0.9% (1/110)
whatsapp_phishing	156	28.8% (45/156)	52	1.9% (1/52)
ransomware	60	8.3% (5/60)	88	1.1% (1/88)
Overall	459	34.6% (159/459)	322	3.1% (10/322)

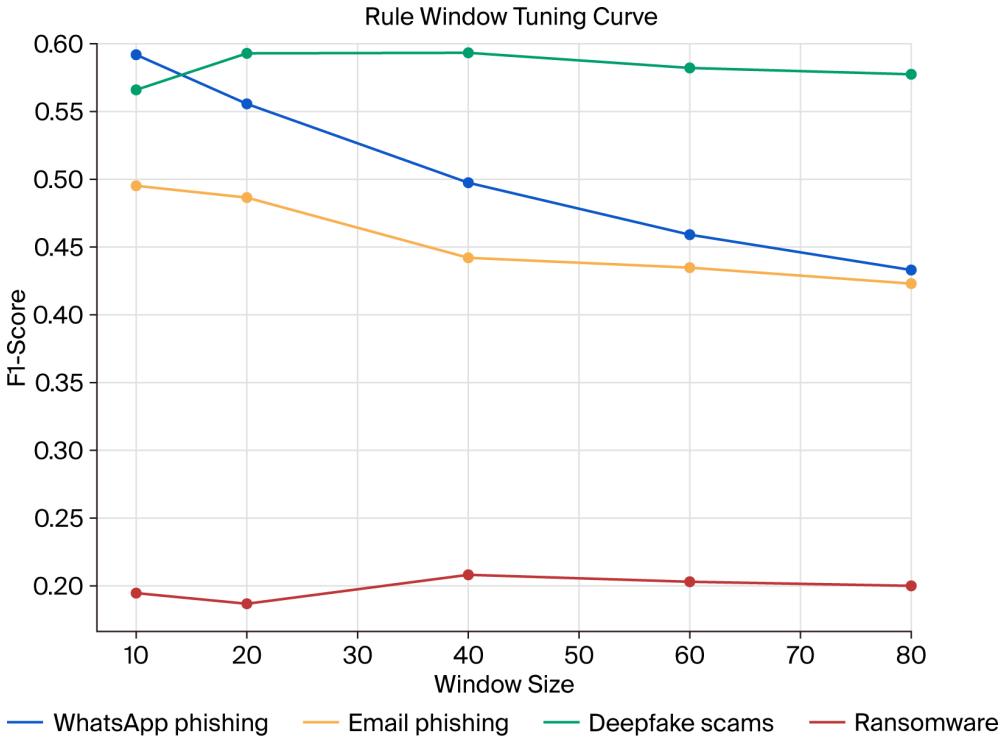


Figure 4. Rule window tuning curve for each ICTT subcategory.

Analysis of the two weak supervision sources reveals significant performance disparities. Regex-based rules achieved a moderate 34.6% overall precision, though results varied from 45.5% for Deepfake Scams and Email Phishing to only 8.3% for Ransomware. The poor performance in the ransomware category stems from semantic ambiguity within Indonesian social media, where general awareness posts and news reports are often structurally indistinguishable from actual incident reports. In contrast, the more distinct linguistic markers of Deepfake Scams and Email Phishing allow for more effective rule-based detection. The file-based labeling approach performed poorly, with an overall precision of just 3.1%. This result highlights a core limitation of keyword-based mining: keywords can ensure topical relevance but cannot confirm that a post describes a specific threat incident. Therefore, the file-based approach is best utilized as a broad initial filter, while regex-based rules offer the granularity required for precise threat-type classification.

Noise in the weakly labeled dataset is classified into two types: Type 1 noise, where the weak label assigns a valid but incorrect threat category, and Type 2 noise, where a threat category is assigned to a post that the expert labeled as "Other." While Type 1 involves categorical misclassification, Type 2 signifies a false positive. Table 4 provides a detailed report of the noise type distribution by source.

The dominant noise type across both sources is Type 2 (false positive), accounting for 63.8% of rule-labeled samples and 96.9% of file-labeled samples. Type 1 noise (cross-category mislabeling) is extremely rare in rule-labeled samples (1.5%, or 7 of 459) and absent

in file-labeled samples (0.0%). This noise profile is structurally favourable for BERT-based fine-tuning: the model is exposed to a large volume of noisy positive examples but is not systematically misled into confusing one threat category with another. This aligns with findings that BERT-based models are robust to label noise, particularly when cross-category mislabeling is rare and the noise is predominantly of the false-positive type [20].

Table 4. Noise type distribution by weak supervision source.

Noise Type	Rule-Labeled (n = 459)	File-Labeled (n = 322)
True Positive (weak label = gold label, gold ≠ "other")	159 (34.6%)	10 (3.1%)
Type 1: Wrong Category (weak ≠ gold, gold ≠ "other")	7 (1.5%)	0 (0.0%)
Type 2: False Positive (gold = "other")	293 (63.8%)	312 (96.9%)
Total	459	322

Of the 1030 rule-labeled samples in the full training set, 459 (44.6%) overlap with the gold-standard set and are directly validated. The remaining 571 rule-labeled samples (55.4%) have no gold-standard counterpart and thus represent unvalidated training data. Similarly, of the 1314 file-labeled samples, 322 (24.5%) overlap with the gold set. While the validated subsets provide reliable estimates of noise characteristics for each source, it is possible that the unvalidated portions exhibit different noise patterns. This limitation is acknowledged in Section 5, and future work should consider expanding the gold-standard annotation or employing noise-aware training techniques to further mitigate label noise impact.

3.4.2. Transformer-Based Classifier (IndoBERT)

At the core of the machine learning approach is the fine-tuned IndoBERT model. This model was trained on the 2344 samples collected (proportioned as shown in Figure 5), using labels generated by the rule-based classifier (weak labeling). The fine-tuning process involved the following:

- Training configuration using the `indolem/indobert-base-uncased` checkpoint with a maximum sequence length of 160 tokens, a batch size of 16, and three training epochs.
- A classification head (a linear layer with softmax activation) was added on top of the IndoBERT base model.
- To mitigate class imbalance, inverse-frequency class weighting was applied via a weighted cross-entropy loss function. This approach avoids the risk of creating linguistically implausible synthetic examples inherent in the Synthetic Minority Over-Sampling Technique (SMOTE) [31].
- The learning rate was set to 5×10^{-5} with weight decay of 0.01, and 10% of the data were reserved as a validation split.
- AdamW optimization and a random seed of 42 were used to ensure reproducible splits and initialization.

The use of a model pre-trained specifically on Indonesian text is critical for understanding local slang, abbreviations, and contextual nuances that would be missed by general multilingual models.

3.4.3. Hybrid Model

The hybrid model combines predictions from both the rule-based classifier and IndoBERT using a threshold-based fusion mechanism. The decision logic operates as follows: if the IndoBERT model's prediction confidence (maximum softmax probability) for a sample exceeds a pre-determined threshold τ^*, the model's prediction is used. If the confidence

falls below τ*, the system falls back to the output of the rule-based classifier, which is more transparent and interpretable. The threshold τ* is optimized separately for each threat subcategory to maximize the F1-score on a validation set. This approach ensures that the system leverages the superior generalization of IndoBERT while preserving the option to fall back on interpretable, auditable rule-based decisions in cases of high uncertainty. The flowchart in Figure 6 illustrates this decision mechanism.

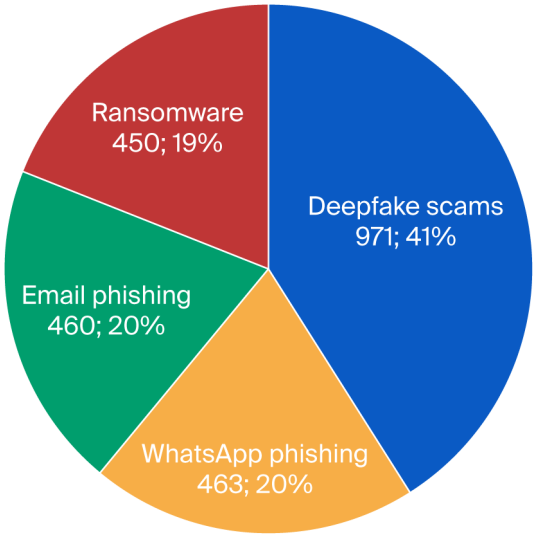


Figure 5. Proportions of weak label distribution.

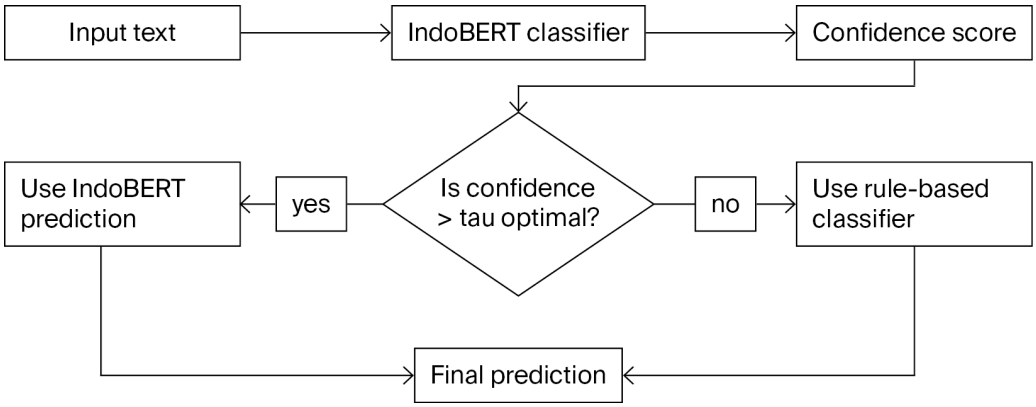


Figure 6. Hybrid model decision logic.

The hybrid model employs a confidence-based decision mechanism. When IndoBERT's confidence exceeds the threshold τ*, its prediction is used. Otherwise, the interpretable rule-based classifier provides the final prediction, ensuring transparency in high-uncertainty cases. Table 5 summarizes the final threshold τ* and F1 per ICTT subcategory.

The fallback frequency was quantified to clarify the operational behavior of the hybrid architecture. This metric represents the proportion of instances in which IndoBERT confidence falls below the class-specific threshold τ*, necessitating invocation of the rule-based classifier. In the 600-sample evaluation set, the fallback mechanism was not triggered, as IndoBERT predictions consistently exceeded the established thresholds. Calibration analysis performed on the 124 ICTT-labeled samples (excluding the "Other" category) yielded an Expected Calibration Error (ECE) of 0.026. This result demonstrates well-aligned confidence estimates. Figure 7 illustrates that most predictions fall within the 0.9 to 1.0 confidence interval, while the reliability diagram in Figure 8 confirms the close alignment between the predicted probability and empirical accuracy.

Table 5. Final threshold τ* and F1 per ICTT subcategory.

Subcategory	τ*	Rule F1	Hybrid F1 (τ*)
deepfake_scams	0.00	0.97	0.99
email_phishing	0.63	0.98	0.98
ransomware	1.00	0.80	0.80
whatsapp_phishing	0.31	0.96	0.96

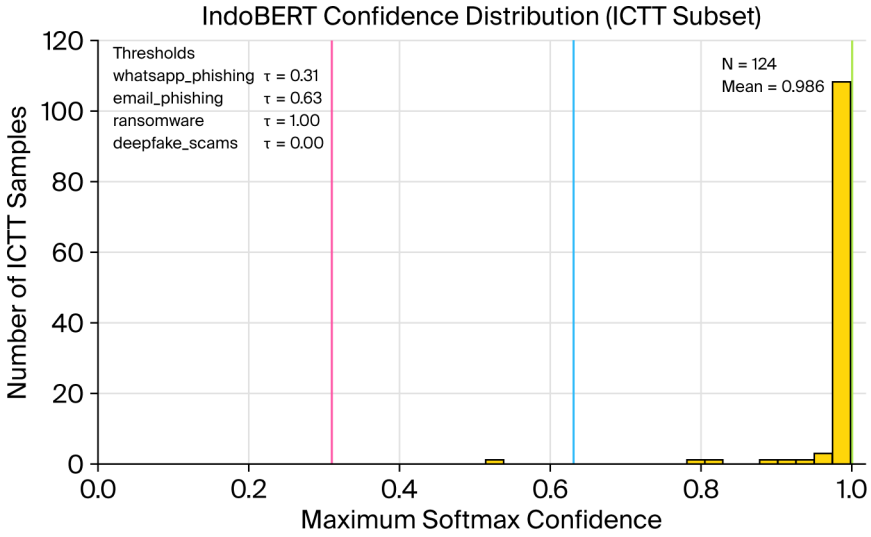


Figure 7. IndoBERT confidence distribution.

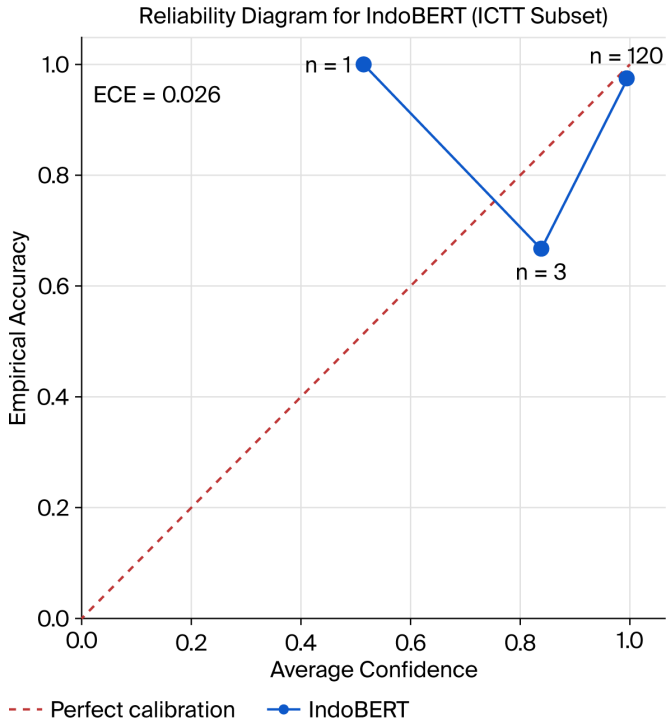


Figure 8. Reliability diagram for IndoBERT (ICTT subset).

Table 6 summarizes the fallback statistics and calibration metrics, and Table 7 reports per-class confidence statistics. These results explain why the hybrid classifier produced identical predictions to IndoBERT during evaluation while still providing a confidence-gated fallback mechanism.

Table 6. Hybrid fallback and calibration summary.

Metric	Value
ICTT evaluation samples	124
IndoBERT accuracy	96.8%
Hybrid fallback count	0
Hybrid fallback rate	0%
Mean confidence	0.986
Median confidence	0.996
ECE (10 bins)	0.026

Table 7. Per-class confidence statistics.

Subcategory	Support	Mean Conf.	Median Conf.
deepfake_scams	39	0.972	0.996
email_phishing	30	0.995	0.996
ransomware	9	0.980	0.996
whatsapp_phishing	46	0.992	0.995

3.5. Evaluation Framework

To rigorously evaluate the models, a gold-standard dataset was created by manually annotating 600 samples with expert-level quality. The annotation of the 600-sample gold-standard dataset was performed by a single domain expert, a deliberate methodological choice justified by the specialized and novel nature of the annotation task and the extreme scarcity of qualified annotators in this domain. The expert holds four professional certifications (CAP, CEH, CRTA, and OSCP+) and possesses hands-on experience in cybersecurity penetration testing and application security assessment. Critically, the expert is fluent in Indonesian and deeply familiar with local cybersecurity discourse, slang, and regional linguistic variations, a rare combination of qualifications in the Indonesian cybersecurity community. Recent research on annotation practices in specialized domains [32] formally recognizes that expert annotators in such fields are "expensive, sometimes inaccessible," and that "the cost of acquiring additional expert annotations is prohibitively high." This principle directly applies to cybersecurity annotation in Indonesian contexts, where qualified annotators capable of distinguishing threat types in informal social media discourse are extremely rare. The single-expert model ensures that all 600 samples are annotated under a unified conceptual framework, which is particularly important for novel constructs such as the ICTT's five dimensions.

Following best practices for specialized-domain annotation [33], annotation was guided by explicit, documented guidelines requiring (1) the assignment of a single dominant ICTT subcategory reflecting the primary threat, (2) decisions based on semantic meaning rather than keyword presence alone, (3) the use of an 'other' category for ambiguous or incomplete content, (4) recording of annotation notes for borderline cases, and (5) avoidance of inferring intent beyond what is reasonably supported by the text. This structured approach, combined with the full traceability of annotation decisions through detailed notes and guidelines, provides transparency and auditability that enables readers to understand and evaluate the annotation rationale. The single-expert design is justified by the scarcity of annotators possessing the rare combination of Indonesian language fluency, cybersecurity expertise, and knowledge of Indonesian regulatory context required for this novel taxonomy.

Model performance was evaluated using the following standard classification metrics:

- Accuracy, the proportion of correctly classified samples, is calculated as

$$\left[\text{Accuracy} = \frac{\sum_{i=1}^{N} I(y_i = \hat{y}_i)}{N}\right]$$

where (I(·)) is the indicator function.

- Precision, Recall, and F1-Score were calculated on a per-class basis to assess performance on individual threat categories.

$$\left[\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}\right]$$

$$\left[\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}\right]$$

$$\left[\text{F1-Score}_c = 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}\right]$$

- Macro-Averaged F1-Score, the unweighted mean of the F1-scores across all classes, provides a balanced view of performance on both majority and minority classes.

$$\left[\text{Micro-F1} = 2 \cdot \frac{\sum_c TP_c}{\sum_c (TP_c + FP_c) + \sum_c (TP_c + FN_c)}\right]$$

4. Results

This section describes the descriptive power of the ICTT and comparative performance of the three classification models on the 600-sample gold-labeled dataset.

4.1. Descriptive Power of ICTT

To empirically evaluate the descriptive power of the ICTT relative to international standards within the context of Indonesian cybercrime discourse, a domain expert retrospectively mapped the 600 gold-standard samples to the ENISA Threat Taxonomy. Each sample was assessed for unambiguous alignment with a single ENISA category, potential ambiguity across multiple categories, or classification entirely outside the taxonomy's scope. Under the ICTT, 124 samples (20.7%) were unambiguously classified. The remaining 476 samples (79.3%) were categorized as "Other," which reflects the expert's stringent annotation criteria. In contrast, only 85 samples (14.2%) could be uniquely mapped to ENISA categories, with 39 (6.5%) identified as ambiguous and 476 (79.3%) falling outside its scope. This represents a 6.5 percentage-point improvement in the unambiguous classification rate for ICTT over ENISA, as summarized in Table 8.

Table 8. Classification clarity: ICTT vs. ENISA (N = 600 gold-standard samples).

Metric	ICTT	ENISA
Unambiguous Classification	124 (20.7%)	85 (14.2%)
Ambiguous/Multiple Mappings	Not available	39 (6.5%)
Rejected/"Other"	476 (79.3%)	476 (79.3%)
Indonesian-Specific Threats Covered	85 samples (68.5% of valid)	Not covered
Deepfake Scams	39 samples (explicit category)	Mapped to Social Engineering (indirect)
WhatsApp Phishing	46 samples (explicit category)	Mapped to Phishing (email-centric; indirect)

The most significant finding is the composition of the ICTT's unambiguous classifications: 85 of the 124 valid ICTT samples (68.5%) belong to Indonesia-specific threat types for which ENISA provides no dedicated category. These include Deepfake Scams (n = 39, 31.5%), WhatsApp Phishing (n = 46, 37.1%), and other platform- or context-specific threats.

Under ENISA, these samples would be mapped indirectly to Social Engineering, Phishing, or Fraud, losing the contextual specificity that distinguishes

- AI-generated voice/video fraud (deepfakes) from conventional social engineering;
- WhatsApp OTP hijacking from email phishing;
- Platform-specific e-wallet fraud (OVO, GoPay, DANA, LinkAja, and QRIS) from generic payment fraud;
- Illegal online loans from conventional fraud.

This loss of granularity is analytically consequential in the Indonesian cybersecurity context. Deepfake scams and WhatsApp phishing are the two predominant emerging cybercrime vectors in Indonesia [3] and require distinct detection methodologies, response protocols, and regulatory approaches. ICTT's explicit categorization of these threats enables more precise threat intelligence, more targeted policy responses, and more effective resource allocation within the Indonesian regulatory framework.

4.2. Overall Model Performance

Table 9 summarizes the overall performance of all four classification models evaluated against the 600 gold-standard samples. The IndoBERT and hybrid models demonstrate a dramatic improvement over both rule-based models, achieving 96.8% accuracy compared to 66.7% for Rule (Tuned) and 61.8% for Rule (Baseline). The Macro-F1 score, which is critical for evaluating performance on imbalanced datasets, shows a similar trend, with the IndoBERT and hybrid models (0.935) far outperforming the Rule (Tuned) model (0.531) and Rule (Baseline) (0.503).

Table 9. Overall comparison (rule-based vs. IndoBERT vs. hybrid).

Model	Accuracy	Macro-Precision	Macro-Recall	Macro-F1	Samples Evaluated
Rule (Baseline)	0.618	0.454	0.765	0.503	600
Rule (Tuned)	0.667	0.473	0.771	0.531	600
IndoBERT	0.968	0.977	0.910	0.935	124 *
Hybrid (rule + IndoBERT)	0.968	0.977	0.910	0.935	124 *

* IndoBERT and Hybrid are evaluated on 124 ICTT-labeled threat samples; 476 samples classified as "other" are excluded from this evaluation subset.

The identical performance of the IndoBERT and Hybrid models masks an important consideration: class imbalance. For WhatsApp Phishing (37.1% of threat-positive samples), F1 = 0.968, while F1 = 0.800 for Ransomware (7.3% of samples). This differential partly reflects training data distribution, suggesting that production systems may exhibit higher sensitivity to WhatsApp Phishing while potentially missing rarer Ransomware incidents.

These results confirm that fine-tuning a pre-trained Indonesian language model on weakly labeled data is a highly effective strategy for this classification task. The hybrid classifier produced the same results in this evaluation. This outcome may seem surprising initially, but it follows naturally from the threshold analysis in the previous section. IndoBERT assigns very high confidence to the correct class for most items, and the learned thresholds τ^* rarely block these predictions. As a result, the hybrid mechanism almost never falls back to the rule-based label.

4.3. Per-Class Performance Analysis

Tables 10 and 11 present the per-class performance of all four models, providing a detailed breakdown that complements the aggregate metrics in Table 9. The four classifiers form a clear progression: the rule-based system succeeds when threat descriptions contain recognizable cues, IndoBERT captures more varied language and interprets posts that

deviate from fixed templates, and the hybrid classifier combines these advantages to reduce the risk of brittle decision-making. Although the difference between IndoBERT and the hybrid system is sometimes small, the hybrid system offers a level of reliability that can be important in operational settings where false alarms and missed incidents carry different costs.

Table 10. Per-class performance—rule-based models (evaluated on 600 gold samples).

Threat Category	Rule (Baseline) Precision/Recall/F1	Rule (Tuned) Precision/Recall/F1	Support
Deepfake Scams	0.470/0.775/0.585	0.463/0.775/0.579	40
Email Phishing	0.420/0.967/0.586	0.431/0.933/0.589	30
Ransomware	0.085/0.556/0.147	0.083/0.556/0.145	9
WhatsApp Phishing	0.333/0.978/0.497	0.425/0.978/0.592	46
Other	0.963/0.549/0.700	0.964/0.613/0.749	475
Macro Avg	0.454/0.765/0.503	0.473/0.771/0.531	600

Table 11. Per-class performance—IndoBERT and hybrid models (evaluated on 124 threat samples).

Threat Category	Precision	Recall	F1-Score	Support
Deepfake Scams	1.000	0.974	0.987	40
Email Phishing	0.968	1.000	0.984	30
Ransomware	1.000	0.667	0.800	9
WhatsApp Phishing	0.939	1.000	0.968	46
Macro Avg	0.977	0.910	0.935	600

The per-class performance gap reflects class imbalance: WhatsApp Phishing (46 samples) achieves F1 = 0.968, while Ransomware (9 samples) achieves F1 = 0.800. This imbalance is analytically consequential and suggests that future work should prioritize collecting additional ransomware samples or exploring techniques such as oversampling to improve performance on rare threat categories.

The performance of the hybrid model is best understood not by evaluating its components in isolation but by recognizing the synergistic relationship between the rule-based system and the IndoBERT model. This hybrid approach, which combines a machine learning algorithm with a rule-based system, has been shown to be effective in text categorization because it leverages the advantages of both [21].

The rule-based classifier, with its high recall and low precision, effectively acts as a high-volume, noisy "threat detector." The IndoBERT model [17] then functions as a sophisticated "threat verifier." Its robustness to the noisy labels generated by the rules is consistent with findings that BERT-based models can be surprisingly robust to label noise, often outperforming more complex noise-handling techniques [20].

This two-stage process, inspired by weak supervision principles [19], is effective because it leverages the speed and transparency of rules for initial filtering and the deep contextual understanding of a transformer for final, high-accuracy classification.

4.4. Platform-Specific Performance

Evaluating classifiers separately across X and YouTube reveals how Indonesian cybercrime discourse varies by platform. While X content is characterized by short, informal, and slang-heavy interactions, YouTube data typically features more structured and explanatory language. The evaluation subset reflects a significant distribution imbalance: out of the 124 ICTT-labeled samples, 109 originate from X, and only 16 come from YouTube. Consequently, performance metrics for YouTube are derived from a much smaller sample

size, which must be considered when interpreting the stability and variety of the results presented in Tables 12 and 13.

Table 12. Platform-specific overall performance.

Model	X Accuracy	X Macro-F1	YouTube Accuracy	YouTube Macro-F1	X Sample	YouTube Samples
Rule (Tuned)	0.982	0.966	0.813	0.803	109	16
IndoBERT	0.991	0.973	0.813	0.803	109	16
Hybrid (Rule + IndoBERT)	0.991	0.973	0.813	0.803	109	16

Table 13. Platform-specific per-class F1-score.

Threat Category	Rule (Tuned) X/YT	IndoBERT X/YT	Hybrid X/YT	X Support	YT Support
Deepfake Scams	0.986/0.889	1.000/0.889	1.000/0.889	35	5
Email Phishing	0.983/1.000	0.983/1.000	0.983/1.000	29	1
Ransomware	0.909/0.500	0.909/0.500	0.909/0.500	6	3
WhatsApp Phishing	0.987/0.824	1.000/0.824	1.000/0.824	39	7

While IndoBERT generalizes across platforms (macro-F1 = 0.973 on X, 0.803 on YouTube), the YouTube evaluation is based on a substantially smaller sample (16 vs. 109 threat-positive samples), limiting statistical confidence in cross-platform claims. Additionally, threat type distribution differs across platforms (e.g., Deepfake Scams: 87.2% on X vs. 12.8% on YouTube), reflecting different user populations and communication styles. The model's performance on emerging platforms (TikTok, Instagram, and Telegram) remains unknown.

The rule-based classifier shows sensitivity to these linguistic differences, performing reasonably well on X's direct cues but struggling with the narrative style of YouTube descriptions that lack specific lexical triggers. In contrast, IndoBERT demonstrates superior generalization, achieving nearly identical macro-F1 scores on both platforms.

By leveraging contextual understanding rather than surface-level keywords, IndoBERT effectively accommodates both compressed X shorthand and longer YouTube sequences. While the limited YouTube sample size necessitates caution, the results suggest that IndoBERT captures underlying threat semantics that are not strictly bound to platform-specific conventions.

A comparative analysis of Tables 12 and 13 confirms the model's capacity to generalize across diverse input lengths, ranging from verbose YouTube metadata to succinct X content. By leveraging deep contextual understanding rather than relying on isolated keywords, the architecture effectively bridges the gap between disparate platform tones and structures, resulting in consistent performance across varying data sources.

4.5. The Role of the 'Other' Category

The 'other' category was used to label content that did not clearly fall into one of the defined threat subcategories. The IndoBERT model achieved a high F1-score of 0.97 for this category, indicating it is highly effective at rejecting out-of-scope or irrelevant content. This capability is particularly important for real-world deployment. A key mode of failure for classifiers is confidently assigning a label to an input that belongs to no defined class. The IndoBERT model's high precision in the 'other' category indicates that it has learned not only what constitutes a "Deepfake Scam" but also what does not, which is

critical for maintaining the reliability of automated threat detection systems in production environments [34,35].

5. Conclusions

Indonesia's digital transformation has generated a vast volume of user-generated cybercrime discourse on social media, yet this intelligence remains largely inaccessible because international taxonomies fail to account for regional linguistic and regulatory nuances. This research addressed this gap by introducing the Indonesian Cybercrime Threat Taxonomy (ICTT), an end-to-end OSINT pipeline, and a hybrid classification architecture. Empirical evaluation against a 600-sample gold-standard dataset confirms the ICTT's superior descriptive power, achieving a 6.5 percentage-point improvement in unambiguous classification over the ENISA framework. Notably, 68.5% of valid classifications involved Indonesian-specific threats, such as WhatsApp phishing and e-wallet fraud, which international standards often overlook.

The hybrid model, combining 12 high-precision regex patterns with a fine-tuned IndoBERT learner, achieved 96.8% accuracy. While IndoBERT provides high-performance contextual understanding, the hybrid architecture adds a critical layer of interpretability. By falling back to rule-based predictions when model confidence is low, the system provides a transparent audit trail for analysts, which is essential in high-stakes cybersecurity environments. Furthermore, this study demonstrates that state-of-the-art performance can be achieved without massive manual annotation. By using weak supervision to programmatically amplify expert knowledge, this approach offers a resource-efficient blueprint for developing localized threat intelligence in other low-resource language contexts globally.

Several methodological decisions were made to ensure the feasibility and internal consistency of the study. Data acquisition was conducted over a three-week window from 19 October to 10 November 2025. This timeframe was a deliberate choice to ensure temporal coherence and a stable snapshot of the threat landscape, avoiding confounding effects from major platform algorithm updates or policy shifts. Additionally, the gold-standard annotation was performed by a single domain expert. This decision reflects the scarcity of specialized cybersecurity expertise in low-resource linguistic contexts and ensures a high degree of internal consistency in applying the ICTT's nuanced classification rules.

Despite these strengths, the study operates within specific boundaries that inform the direction of future work. The rule-based component is limited to 12 patterns, and the training data relies on weak supervision that currently covers a subset of the total mined samples. Validation indicates that rule-based labels achieve 34.6% precision, yet this covers only 44.6% of the rule-labeled samples and 24.5% of the file-labeled samples. Furthermore, representational biases, including a class imbalance (a 5.1:1 ratio between WhatsApp Phishing and Ransomware), platform skew (75.7% X versus 24.3% YouTube), and user concentration where the top user represents 6% of the data, require careful interpretation. Although these biases reflect the real threat landscape and communication patterns on Indonesian social media rather than poor collection methods, they emphasize the need for ongoing model calibration to ensure fairness across threat types and user groups.

Future work should transition from this controlled snapshot to a longitudinal analysis by extending the collection period across multiple seasons and events to reduce temporal bias. Incorporating multiple domain experts to assess inter-rater reliability will be essential for validating the taxonomy's objectivity and mitigate potential annotation bias. Natural extensions also include stratified sampling across user types to improve model fairness, user deduplication to reduce individual influence, and the exploration of noise-aware training techniques, such as mixup, to enhance resilience. In summary, the ICTT and its accompanying hybrid system establish a scalable, interpretable foundation for cyber

threat analytics in Indonesia, supporting the operational needs of stakeholders like the Indonesian Cybersecurity Agency (BSSN) while providing a replicable framework for the international community.

Author Contributions: Conceptualization, F.A. and T.M.; methodology, F.A. and T.M.; software, F.A.; validation, T.M., D.O.D.H. and F.A.; formal analysis, F.A. and D.O.D.H.; investigation, D.O.D.H.; resources, F.A.; data curation, D.O.D.H.; writing—original draft preparation, F.A.; writing—review and editing, T.M. and D.O.D.H.; supervision, T.M. and D.O.D.H.; project administration, F.A.; funding acquisition, F.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is available upon request from the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. APJII. *Survei Penetrasi dan Perilaku Internet Indonesia 2025 [Internet Penetration and Behavior Survey of Indonesia 2025]*; Asosiasi Penyelenggara Jasa Internet Indonesia: Jakarta, Indonesia, 2025.
2. Dhanya, D. Indonesia's BSSN Records 3.64 Billion Cyberattacks in First Half of 2025. Tempo English. 2025. Available online: https://en.tempo.co/read/2037469/indonesias-bssn-records-3-64-billion-cyberattacks-in-first-half-of-2025 (accessed on 22 September 2025).
3. BSSN. *Lanskap Keamanan Siber Indonesia 2024*; Badan Siber dan Sandi Negara (BSSN): Jakarta, Indonesia, 2024. Available online: https://www.scribd.com/document/834167154/LANSKAP-KEAMANAN-SIBER-2024 (accessed on 22 September 2025).
4. OJK. *OJK Performance Report 2024*; OJK: Jakarta, Indonesia, 2024.
5. Marinos, L. ENISA Threat Taxonomy A Tool for Structuring Threat Information. Heraklion. 2016. Available online: https://www.enisa.europa.eu (accessed on 22 September 2025).
6. Europol. *Internet Organised Crime Threat Assessment (IOCTA) 2023*; Publications Office of the European Union: Luxembourg, 2023. [CrossRef]
7. Verizon. *DBIR 2023 Data Breach Investigations Report*; Verizon: New York, NY, USA, 2023.
8. MITRE Corporation. *MITRE ATT&CK: A Knowledge Base of Adversary Tactics and Techniques*; MITRE Corporation: McLean, VA, USA, 2018. Available online: https://attack.mitre.org/ (accessed on 20 December 2025).
9. Chandra, A.; Snowe, M.J. A taxonomy of cybercrime: Theory and design. *Int. J. Account. Inf. Syst.* **2020**, *38*, 100467. [CrossRef]
10. Agrafiotis, I.; Nurse, J.R.C.; Goldsmith, M.; Creese, S.; Upton, D. A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate. *J. Cybersecur.* **2018**, *4*, tyy006. [CrossRef]
11. Malavasi, M.; Peters, G.W.; Trück, S.; Shevchenko, P.V.; Jang, J.; Sofronov, G. Cyber risk taxonomies: Statistical analysis of cybersecurity risk classifications. *Insur. Math. Econ.* **2026**, *126*, 103167. [CrossRef]
12. Arazzi, M.; Arikkat, D.R.; Nicolazzo, S.; Nocera, A.; K.A., R.R.; P., V.; Conti, M. NLP-based techniques for cyber threat intelligence. *Comput. Sci. Rev.* **2025**, *58*, 100765. [CrossRef]
13. Albarrak, M.; Salonitis, K.; Jagtap, S. Natural language processing (NLP)-based frameworks for cyber threat intelligence and early prediction of cyberattacks in Industry 4.0: A systematic literature review. *Appl. Sci.* **2026**, *16*, 619. [CrossRef]
14. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**. [CrossRef]
15. Bayer, M.; Kuehn, P.; Shanehsaz, R.; Reuter, C. CySecBERT: A domain-adapted language model for the cybersecurity domain. *ACM Trans. Priv. Secur.* **2024**, *27*, 18. [CrossRef]
16. Aghaei, E.; Niu, X.; Shadid, W.; Al-Shaer, E. SecureBERT: A domain-specific language model for cybersecurity. In *Security and Privacy in Communication Networks (SecureComm 2022)*; Li, F., Liang, K., Lin, Z., Katsikas, S.K., Eds.; Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering; Springer: Cham, Switzerland, 2023; Volume 462. [CrossRef]
17. Koto, F.; Rahimi, A.; Lau, J.H.; Baldwin, T. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 757–770. [CrossRef]

18. Wilie, B.; Vincentio, K.; Winata, G.I.; Cahyawijaya, S.; Li, X.; Lim, Z.Y.; Soleman, S.; Mahendra, R.; Fung, P.; Bahar, S.; et al. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Suzhou, China, 4–7 December 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 843–857. [CrossRef]
19. Ratner, A.; Bach, S.H.; Ehrenberg, H.; Fries, J.; Wu, S.; Ré, C. Snorkel: Rapid Training Data Creation with Weak Supervision. *Proc. VLDB Endow.* **2017**, *11*, 269–282. [CrossRef] [PubMed]
20. Zhu, D.; Hedderich, M.A.; Zhai, F.; Adelani, D.I.; Klakow, D. Is BERT robust to label noise? A study on learning with noisy labels in text classification. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP, Dublin, Ireland, 26 May 2022*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 62–67. [CrossRef]
21. Villena-Román, J.; Collada-Pérez, S.; Lana-Serrano, S.; González, J.C. Hybrid approach combining machine learning and a rule-based expert system for text categorization. In Proceedings of the Florida Artificial Intelligence Research Society Conference, (FLAIRS), Palm Beach, FL, USA, 18–20 May 2011.
22. Li, X.; Cui, M.; Li, J.; Bai, R.; Lu, Z.; Aickelin, U. A hybrid medical text classification framework: Integrating attentive rule construction and neural network. *Neurocomputing* **2021**, *443*, 345–355. [CrossRef]
23. Samtani, S.; Chinn, R.; Chen, H.; Nunamaker, J.F., Jr. Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence. *J. Manag. Inf. Syst.* **2017**, *34*, 1023–1053. [CrossRef]
24. Shaukat, K.; Luo, S.; Chen, S.; Liu, D. Cyber Threat Detection Using Machine Learning Techniques: A Performance Evaluation Perspective. In *Proceedings of the 1st Annual International Conference on Cyber Warfare and Security, Virtual, 20–21 October 2020*; ICCWS 2020-Proceedings; IEEE: New York, NY, USA, 2020. [CrossRef]
25. Sarabi, A.; Huang, Z.; Wang, C.; Karir, T.; Liu, M. The Ransomware Decade: The Creation of a Fine-Grained Dataset and a Longitudinal Study. In Proceedings of the 34th USENIX Security Symposium, Seattle, WA, USA, 13–15 August 2025. [CrossRef]
26. Demirol, D.; Das, R.; Hanbay, D. A Novel Approach for Cyber Threat Analysis Systems Using BERT Model from Cyber Threat Intelligence Data. *Symmetry* **2025**, *17*, 587. [CrossRef]
27. Wirth, R.; Hipp, J. CRISP-DM: Towards a Standard Process Model for Data Mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, 11–13 April 2000; pp. 29–40.
28. Schröer, C.; Kruse, F.; Gómez, J.M. A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Comput. Sci.* **2021**, *181*, 526–534. [CrossRef]
29. Martínez-Plumed, F.; Contreras-Ochando, L.; Ferri, C.; Hernández-Orallo, J.; Kull, M.; Lachiche, N.; Ramírez-Quintana, M.J.; Flach, P. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Trans. Knowl. Data Eng.* **2019**, *33*, 3048–3061. [CrossRef]
30. Achuthan, K.; Khobragade, S.; Kowalski, R. Cybercrime through the public lens: A longitudinal analysis. *Humanit. Soc. Sci. Commun.* **2025**, *12*, 282. [CrossRef]
31. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
32. Calderon, N.; Reichart, R.; Dror, R. The Alternative Annotator Test for LLM-as-a-Judge: How to Statistically Justify Replacing Human Annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025), Vienna, Austria, 27 July–1 August 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; Volume 1, pp. 16051–16081.
33. Beck, J. Quality aspects of annotated data: A research synthesis. *AStA Wirtsch. Sozialstatistisches Arch.* **2023**, *17*, 331–353. [CrossRef]
34. Al-garadi, M.A.; Varathan, K.D.; Ravana, S.D. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Comput. Hum. Behav.* **2016**, *63*, 433–443. [CrossRef]
35. Karteris, A.; Tzanos, G.; Papadopoulos, L.; Soudris, D. Detection of Cyber Security Threats through Social Media Platforms. In *2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), St. Petersburg, FL, USA, 15–19 May 2023*; IEEE: New York, NY, USA, 2023; pp. 820–823. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.