MOBILE AND FORENSICS

Vol. 8, No. 2, 2026, pp. 110-126 | DOI: 10.12928/mf.v8i2.16449

ISSN: 2714-6685 (Online)
ISSN: 2656-6257 (Print)

A Hybrid Machine Learning and Digital Forensics Framework for Detecting Coordinated Suspicious Accounts on X

Riko Iman Decamarta [a,1,*], **Yudi Prayudi** [a,2]

[a] Universitas Islam Indonesia, Yogyakarta, Indonesia
[1] 22917016@students.uii.ac.id*; [2] prayudi@uii.ac.id

* Corresponding Author

ARTICLE INFO

ARTICLE HISTORY

Received: Mar 29, 2026
Revised: Jul 10, 2026
Accepted: Jul 19, 2026

KEYWORDS

Fake Accounts, Machine Learning, Stylometry, Digital Forensics, Coordinated Inauthentic Behavior

ABSTRACT

Coordinated Inauthentic Behavior (CIB) on X threatens public discourse, yet single-method approaches fail to detect implicit coordination. This study proposes a conditional hybrid framework spanning four forensic stages: (1) automated classification using multilingual BERT text embeddings concatenated with 21 numerical features, (2) sockpuppet candidate generation via Sentence-BERT stylometric fingerprinting and cosine similarity, (3) graph-based coordination analysis using the Louvain algorithm, and (4) manual forensic confirmation with OSINT investigation. Trained on the Cresci-2017 benchmark which exhibits substantial temporal and linguistic domain mismatch from our 2025 Indonesian target domain, the BERT model achieved an accuracy, precision, recall, and F1-score of 1.00 on its held-out test partition prior to deployment. The framework was applied to 1,309 Indonesian Military Law (UU TNI) tweets from 720 accounts, which resulted in the identification of 114 suspicious accounts (15.8%) and the extraction of 28 candidate accounts from 6 groups. Manual validation by the authors through side-by-side timeline inspection confirmed one pair of accounts with synchronized and verbatim identical content. The absence of direct network interactions despite high stylometric similarity is interpreted as consistent with implicit CIB evasion strategies, but attribution to organized groups cannot be definitively established on the basis of platform data alone.

1. INTRODUCTION

The X (Twitter) platform is one of the most active digital communication platforms in the world, used by millions of people to share information and opinions regarding public issues in real time. However, the openness of this platform also makes it vulnerable to misuse by fake accounts and bot networks that operate in a coordinated manner to manipulate public narratives [1][2]. Fake accounts are defined as fictitious digital identities created for manipulative purposes such as spreading disinformation, strengthening certain narratives, or creating the illusion of massive public support

[1]. Varol et al. estimated that 9 to 15% of active accounts on Twitter are bots operating automatically [3]. Networks of fake accounts can also operate together, a phenomenon known as Coordinated Inauthentic Behavior (CIB), which aims to mislead users and strategically influence public discussion [4][5]. CIB can manifest in two forms: explicit coordination, in which accounts interact directly with one another through mentions, replies, or retweets and therefore form visible edges in a social graph; and implicit coordination, in which accounts act in coordination by publishing highly similar content with consistent linguistic styles and synchronized temporal patterns without any direct interaction, thereby leaving no detectable structural traces in the interaction graph [6][7].

It is important to emphasize that the CIB label carries a strong evidentiary burden, as it implies coordinated deception regarding an actor's identity, origin, or purpose [4]. Textual similarity and synchronized posting alone are insufficient, since such patterns may equally arise from legitimate causes including content syndication, institutional publishing, sanctioned campaign coordination, automated scheduling tools, or cross-account amplification by the same organization [7]. In this research, findings are labeled as consistent with CIB only when a minimum combination of evidence is satisfied: (a) indications of account inauthenticity from automated classification, (b) stylometric similarity exceeding the level attributable to shared topics, (c) statistically improbable repeated temporal synchronization, and (d) the absence of a verifiable legitimate explanation such as an authentic institutional identity or content syndication. The elimination of legitimate explanations under criterion (d) is carried out through manual validation and OSINT investigations. Any finding that fails this criterion is treated as a false positive rather than as evidence of inauthentic coordination.

Previous research on suspicious account detection has generally used single-method approaches based on machine learning [8], [9], account metadata [3], or graph analysis [10]. Single-method approaches have a fundamental limitation: they are unable to identify implicitly coordinated accounts as defined above, since such accounts share similar language styles and temporal patterns but deliberately avoid direct interaction to evade network-based detection [7]. Cresci et al. introduced the concept of social spambots that operate in a sophisticated, human-like manner, making them difficult to distinguish from genuine users even through manual inspection [11].

Hybrid detection itself is not new, and several strands of prior work combine multiple signals. Feature-based systems such as that of Varol et al. aggregate profile, content, and network features into a single per-account classifier [3]; Sahoo and Gupta combine machine learning with profile analysis for malicious profile detection [12]; and Mohammadrezaei et al. pair graph metrics with classification algorithms [10]. These systems, however, operate at the level of individual accounts and do not detect coordination among them. Conversely, coordination-oriented approaches such as RTbust [13] and the coordinated behavior detection methods surveyed and applied by Nizzoli et al., Pacheco et al., and others [7][14][15][16], uncover groups of accounts acting in coordination, but they do not classify account inauthenticity and stop short of forensic evidence reporting, while stylometric sockpuppet detection has largely been studied on long-form platforms such as Wikipedia [17].

Based on Table 1 positions the proposed framework against these strands. The novelty of this research therefore lies not in hybridization per se but in three specific aspects: (1) conditional execution across four forensic stages, in which each stage gates the input of the next so that expensive analyses are applied only to accounts that pass earlier thresholds; (2) the detection of implicit coordination through combined semantic and stylometric similarity of short-text linguistic fingerprints, complemented by temporal synchronization analysis; and (3) the integration of account classification with OSINT investigation into a structured, investigation-ready forensic report, rather than a standalone detection score.

Table 1. Comparison with Prior Hybrid and Coordination Detection Approaches

Approach	Account Classification	Coordination Detection	Stylometric Similarity	Temporal Analysis	Conditional Staging	Forensic OSINT Report
Feature-based bot detection, Varol et al. [3]	Yes	No	No	Partial	No	No

Hybrid ML profile detection, Sahoo and Gupta [12]	Yes	No	No	No	No	No
Graph plus classification, Mohammadrezaei et al. [10]	Yes	No	No	No	No	No
RTbust temporal botnet detection, Mazza et al. [13]	No	Yes	No	Yes	No	No
Coordinated behavior detection, Nizzoli et al.; Cinus et al. [7], [14], [15].	No	Yes	Partial	Yes	No	No
Stylometric sockpuppet detection, Solorio et al. [17]	No	Partial	Yes	No	No	No
Proposed Framework	Yes	Yes	Yes	Yes	Yes	Yes

This research addresses the identified methodological gap by proposing a hybrid framework that integrates BERT-based classification, combining tweet-text embeddings with account-level numerical features [18], stylometric linguistic fingerprinting using Sentence-BERT [6], social graph analysis with Louvain community detection [19], and OSINT investigation [20]. The framework is tested on a dataset of 1,309 tweets from 720 accounts related to the Indonesian Military Law (UU TNI) issue on the X platform in April 2025. The revision of the UU TNI in early 2025 sparked a polarized nationwide debate over the expansion of military roles in civilian affairs and triggered large-scale public protests, creating a high-stakes political controversy in which competing actors have strong incentives to manipulate public opinion, precisely the conditions under which CIB campaigns typically emerge.

This research is expected to contribute a conditional hybrid framework capable of detecting both individual suspicious accounts and coordinated sockpuppet candidate networks, empirical evidence of patterns consistent with implicit CIB coordination in Indonesian public issues, and a valid evidentiary method for the digital forensics of coordinated account networks.

2. METHOD

The proposed method is a conditional hybrid framework consisting of four integrated components, as illustrated in Fig. 1. The conditional design ensures that each account only proceeds to the next stage if it meets the threshold of the previous stage. This approach maintains computational efficiency without sacrificing detection accuracy. The framework produces a digital forensic report as its final output, encompassing a list of individual suspicious accounts, sockpuppet groups, network coordination maps, and OSINT findings.

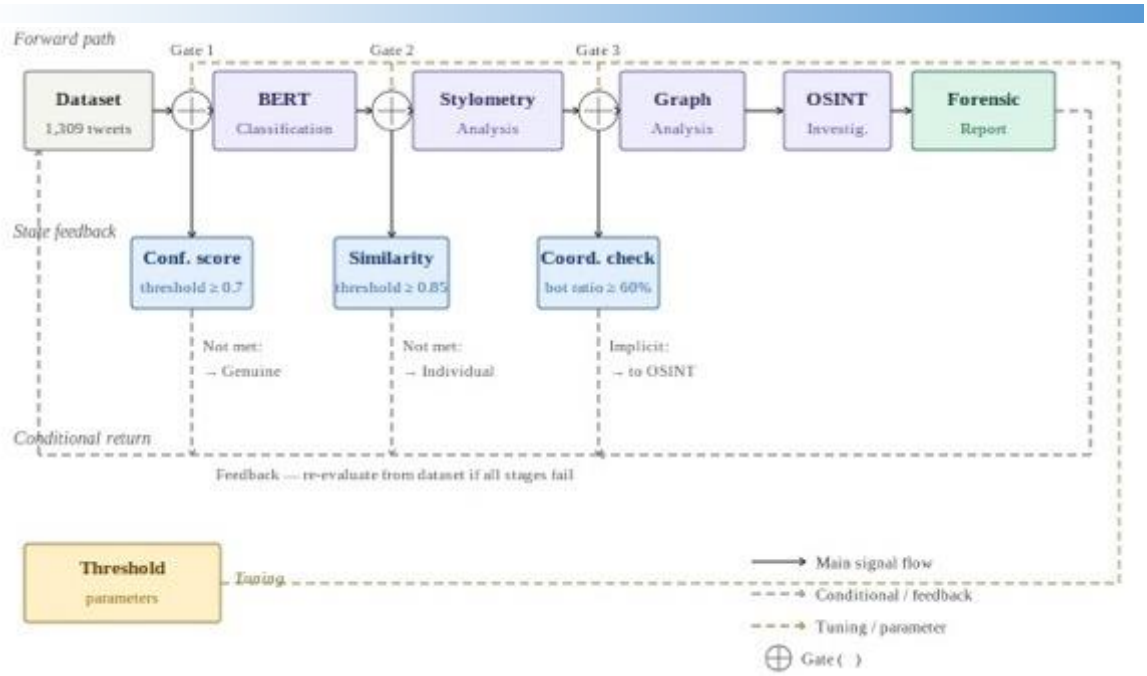

Fig. 1. The proposed method

2.1. Dataset

The case dataset was collected between 11 and 17 April 2025 using the tweet-harvest tool [21], a Node is-based crawler operating on an authenticated X user session, executed on Google Colab with the search query "UU TNI" restricted to Indonesian-language tweets (lang: id). The inclusion criteria were: tweets matching the query keyword, written in Indonesian, and posted within the collection window; both original tweets and replies were included. Pure retweets were excluded by the collection tool's default behavior, which is appropriate for this study, since retweeted text does not represent the retweeting account's authoring style. Content deleted before the collection time cannot be captured, and the collection is limited by the platform search index's coverage; these sampling limitations are acknowledged. The dataset contains 1,309 tweets from 720 unique accounts with 15 attribute columns, including the full tweet text, engagement metrics (retweet count, favorite count, reply count, quote count), and interaction relations (mention and reply targets). Table 2 summarizes the dataset characteristics.

Table 2. Research Dataset Characteristic

Attribute	Value
Total tweets (raw)	1,309 tweets
Unique accounts	720 accounts
Collection period	April 2025
Platform	X (Twitter)
Topic	Indonesian Military Law (UU TNI) issue
Available columns	conversation_id_str, created_at, favorite_count, full_text, id_str, image_url, in_reply_to_screen_name, lang, location, quote_count, reply_count, retweet_count, tweet_url, user_id_str, username

Tweet volume per account is highly skewed: 84.4% of accounts contributed a single tweet (median 1, mean 1.81, maximum 116). No minimum-text criterion was imposed at the classification stage. However, the stylometric fingerprint of single-tweet accounts is inherently weak. This limitation is mitigated by the conservative similarity threshold described below and by the requirement of downstream corroboration before any coordination interpretation is applied.

The BERT training dataset uses Cresci-2017 [11], the most widely cited Twitter bot detection benchmark, containing 500 genuine accounts and 500 social spambot accounts annotated through a crowdsourcing process. A notable limitation is domain mismatch. Cresci-2017 contains Italian-language tweets from 2017, whereas the research data contain Indonesian tweets from 2025, and bot classifiers are known to generalize poorly across domains [22]. Constructing a manually labeled Indonesian ground-truth dataset for direct target-domain validation and applying domain adaptation techniques were beyond the scope of this study and are identified as priority future work. In this study, the mismatch is mitigated architecturally, with stylometric analysis, graph analysis, and manual validation serving as complementary validation layers [23].

2.2. Data Preprocessing

Preprocessing is performed to clean the raw dataset and extract structured features for each account. The preprocessing pipeline consists of 3 sub-stages:

- Text cleaning: removal of URLs (http/https patterns), mention tags (@username), hashtag symbols (# removed, word preserved), emojis, and non-alphanumeric special characters using regular expressions, followed by whitespace normalization.
- Duplicate removal: tweets are deduplicated based on two criteria: identical id_str values and identical combinations of user_id_str and full_text. This step removes 6 duplicate tweets, yielding 1,303 clean tweets.
- Feature extraction: 21 features are extracted per account across three categories and normalized using StandardScaler [24], as detailed in Table 3.

Table 3. Extracted Feature

Feature Group	Features	Count
Engagement & Activity	tweet_count, avg_retweet, avg_favorite, avg_reply, avg_quote, retweet_ratio, reply_ratio, has_location, posting_hour_std	9
Stylometric	avg_tweet_length, type_token_ratio, capital_ratio, exclamation_freq, question_freq, hashtag_rate, url_ratio	7
Social Relations	mention_count, mention_diversity, replied_to_count, replied_to_diversity, interaction_diversity	5
	Total	21

2.3. BERT Classification

The classification model is built upon the bert-base-multilingual-cased pre-trained model [18], which supports 104 languages, including Bahasa Indonesia [25]. The model input is constructed per account. All tweets of an account are concatenated into a single document, tokenized with a maximum sequence length of 128 tokens, and truncation is applied to longer documents. The 768-dimensional [CLS] embedding produced by the BERT encoder is concatenated with the 21 normalized account-level numerical features, and the resulting 789-dimensional vector is passed to a fully connected classification head producing a two-class output. Accounts with little or no textual content after preprocessing are represented by a placeholder text and rely predominantly on their numerical features.

Training uses an account-level 80/10/10 stratified train, validation, and test split with a fixed random seed, applied after each account is aggregated into a single sample, so that no account contributes data to more than one partition. Model selection is based solely on validation loss, and the test partition is evaluated exactly once. One deviation from strict isolation is disclosed. The feature scaler was fitted on the full feature matrix prior to partitioning, which exposes distributional statistics but no label information. An ablation experiment, in which artifact-correlated features were removed and empty-text accounts excluded, was additionally conducted to dissect the benchmark evaluation results (see Results and Discussion).

Accounts with a confidence score ≥ 0.7 are classified as suspicious and forwarded to the semantic-stylometric stage. This score acts as a gating threshold. A lower threshold would pass too many false positives to subsequent stages, while a higher threshold may risk missing genuinely suspicious accounts. Because the benchmark evaluation exhibits perfect class separation, ROC-based threshold optimization is uninformative. The gating threshold is instead assessed through a sensitivity analysis on the target data, reported in the Results section.

2.4. Semantic-Stylometric Analysis

This stage identifies whether suspicious accounts are likely operated by the same entity based on the combined semantic and stylistic similarity of their writing, hereafter referred to as a semantic-stylometric fingerprint. It is acknowledged that Sentence-BERT embeddings primarily capture semantic content rather than authorship style. The fingerprint therefore combines both signal types, on the premise that an operator maintaining multiple accounts will exhibit consistent patterns across those accounts [17][26].

For each account, all tweets are combined into a single document and processed using the paraphrase-multilingual-MiniLM-L12-v2 Sentence-BERT model [6] to produce a 384-dimensional semantic embedding. This embedding is combined with 7 normalized numerical stylometric features, including average tweet length, type-token ratio, capitalization ratio, punctuation frequency, hashtag rate, and URL ratio, yielding a 391-dimensional vector that serves as the fingerprint of each account. The embedding is L2-normalized, and the numerical features are standardized before concatenation. It is acknowledged that the semantic component dominates the representation in terms of dimensionality (384 versus 7). The stylometric features serve as secondary discriminators, and cases where semantic similarity alone yields spurious matches are handled by the downstream validation layers, as demonstrated by the false positive identified in Group 6.

Cosine similarity is computed between all pairs of fingerprint vectors. Pairs with similarity ≥ 0.85 are identified as sockpuppet candidates. This threshold was validated empirically against the distribution of all 6,441 pairwise similarities among the 114 suspicious accounts, which has a median of 0.30 and a 99th percentile of 0.798. Because every account pair in the corpus discusses the same topic, this distribution constitutes the topic-sharing baseline, and the 0.85 threshold lies above its 99th percentile, filtering far beyond mere topical similarity. A threshold sensitivity analysis is reported in the Results section. Candidate pairs are then grouped into candidate clusters using the Union-Find algorithm, in which accounts A, B, and C form a single group if sim(A,B) ≥ 0.85 and sim(B,C) ≥ 0.85, regardless of sim (A,C).

Since Union-Find groups accounts transitively, large components are susceptible to chaining effects, in which accounts are merged through intermediate members despite falling below the pairwise threshold. Two safeguards are applied to characterize and mitigate this risk. First, the internal cohesion of each resulting group is quantified by computing the average and minimum intra-group pairwise similarity, and a complete-linkage decomposition, requiring every internal pair to satisfy the ≥ 0.85 threshold, is applied to identify tightly connected cores within large components. The grouping is additionally compared against average-linkage, complete-linkage, and DBSCAN clustering on the same similarity matrix (see Results). Second, the semantic-stylometric grouping serves only as a candidate-generation stage rather than a final verdict: all group memberships are subsequently cross-validated through social graph analysis, temporal synchronization analysis, and OSINT investigation, so that weakly chained members lacking independent evidence of coordination are assigned lower confidence in the final forensic report.

2.5. Social Graph Analysis

A directed weighted graph G = (V, E, W) is constructed using NetworkX [27]. The 1516 raw interactions extracted during preprocessing are aggregated into unique directed edges, in which repeated interactions between the same ordered account pair are merged and their frequency recorded as the edge weight. The node set exceeds the 720 collected accounts because mention and reply targets include external accounts that did not author tweets in the corpus; of the 720 collected accounts, 436 participated in at least one interaction, and the remaining nodes are interaction targets outside the collection set. Community detection is performed on the undirected projection of this graph using the Louvain algorithm [19] with a fixed random seed for reproducibility, iteratively reassigning nodes to communities to maximize modularity. The topological metrics computed include degree centrality, betweenness centrality, and clustering coefficient.

A community cluster is classified as coordinated if it satisfies at least two of the following three criteria: (1) the bot account ratio within the community is ≥ 60%; (2) the average clustering coefficient of the community exceeds the global average; or (3) the community contains hub nodes whose betweenness centrality exceeds twice the global average, indicating broker or coordinator accounts. The 60% bot-ratio threshold was chosen as a conservative majority-based heuristic: it requires a community to be clearly dominated by suspicious accounts, providing a safety margin above a simple majority (50%) to absorb potential residual misclassification from the BERT stage. Because a community must satisfy at least two of the three criteria, the classification is not sensitive to the exact value of this threshold; a sensitivity analysis varying the bot-ratio threshold from 50% to 70% is reported in the Results section. External validation of these criteria against independently confirmed coordinated and organic communities was not possible in the absence of Indonesian ground-truth data and is identified as a future work item. If no coordinated cluster is found, the coordination is classified as implicit, meaning that candidate accounts do not directly interact with each other on the platform.

2.6. OSINT Investigation

The OSINT investigation is conducted to provide cross-platform evidence and to complement the computational analyses. It comprises three automated sub-components applied to the 28 identified candidate accounts:

- Username lookup: HTTP requests are sent to five platforms (Instagram, TikTok, YouTube, GitHub, and LinkedIn) to detect cross-platform presence under identical usernames. It is noted that matching usernames do not prove common ownership, and the absence of matches is not evidence of evasion, since username reuse across platforms is neither necessary nor sufficient to support either conclusion; lookup results serve only as supplementary investigative leads.
- URL and domain analysis: all URLs shared by candidate accounts are extracted and analyzed using the tldextract library to identify shared external domains. Accounts sharing the same non-platform domain (excluding t.co) are considered to potentially share digital infrastructure.
- Temporal posting pattern analysis: the timestamps of candidate account posts are analyzed to detect synchronization slots, defined as time windows at one-minute granularity in which more than one candidate account posts simultaneously. Because same-minute coincidences can arise organically in an event-driven corpus, the observed slot count is tested against a permutation-based null model (10,000 iterations) that preserves each account's tweet volume and draws timestamps from the empirical distribution of the full corpus, thereby controlling for activity levels, peak hours, and event bursts. The evidentiary interpretation of synchronization is based on this null comparison and on pair-level analysis of which accounts co-post, rather than on the raw slot count.

The output of the OSINT investigation is a structured forensic report documenting all identified synchronization slots, shared domains, and cross-platform lookup results, accompanied by a recommended follow-up investigation procedure in Maltego, including seed entities and suggested transforms to uncover the infrastructure and entities behind the candidate network.

2.7. Manual Validation Protocol

Manual validation was performed by the first author through side-by-side inspection of candidate accounts' full tweet timelines, comparing tweet texts verbatim, posting timestamps, hashtag usage and placement, profile attributes, and account verification status; full chain-of-custody procedures as specified in digital evidence handling standards [28], [29] were not implemented and are recommended for operational deployment. Evidence was preserved as screenshots taken during the inspection, as shown in the article figures. Since validation was conducted by a single annotator, inter-annotator reliability could not be assessed; this is acknowledged as a limitation, partially mitigated by the fact that the confirmed findings rest on objectively verifiable properties, namely verbatim text identity and recorded timestamps, rather than on subjective judgment.

3. RESULTS AND DISCUSSION

3.1. Preprocessing Results

Preprocessing of 1,309 raw tweets produced 1,303 clean tweets (6 duplicates removed), 21 normalized features per account, and 1,516 social relation edges extracted from mention and reply's interactions. These edges serve as the structural input for the social graph analysis stage.

3.2. BERT Classification Results

The BERT model was trained for 10 full epochs, with training loss decreasing from 0.5906 to 0.0043 and validation loss from 0.4113 to 0.0034, indicating no overfitting to the training partition. Evaluation on 100 held-out test accounts from Cresci-2017 achieved Accuracy = Precision = Recall = F1-Score = 1.00 (50 genuine and 50 spambot accounts, all correctly classified; the confusion matrix contains no off-diagonal entries). Given that perfect scores on social spambot benchmarks frequently signal data leakage, the evaluation protocol was verified as described in the Methods; a critical dissection of this score, including artifact analysis and an ablation experiment, is presented in the Framework Discussion.

Based on Table 4, applied to the 720 research accounts, 114 accounts (15.8%) exceeded the 0.70 gating threshold and were forwarded to the semantic-stylometric stage. The gating outcome is robust to the threshold choice: the confidence distribution is strongly bimodal (median confidence of flagged accounts 0.995), yielding 116, 114, 114, 111, and 107 gated accounts at thresholds of 0.60, 0.70, 0.80, 0.90, and 0.95, respectively. It must be emphasized that, in the absence of ground-truth labels for the 720 target accounts, this figure denotes the volume of investigative candidates, not a measured detection accuracy; end-to-end accuracy on the target domain cannot be established, and all downstream findings carry candidate status unless individually validated.

Table 4. BERT Classification Results on 720 Research Accounts

Category	Count	Percentage	Next Action
Classified as genuine	602 accounts	83.6%	Analysis terminated
Classified as fake	118 accounts	16.4%	Further processing
Suspicious (conf ≥ 0.7)	114 accounts	15.8%	Proceed to stylometry

3.3. Semantic-Stylometric Analysis Results

From 6,441 pairwise comparisons among the 114 accounts, 34 pairs exceeded the similarity threshold of ≥ 0.85, involving 28 unique accounts in 6 candidate groups; three pairs showed near-identical similarity (> 0.95). The threshold sensitivity analysis specified in the Methods shows the six-group structure is stable across thresholds from 0.80 to 0.875 (28 ± 3 candidate accounts), with fragmentation occurring only at 0.90, and all tightly matched pairs persisting at every value. Comparison across clustering methods yields the same conclusion: single-linkage (Union-Find), DBSCAN, average-linkage, and complete-linkage recover a nearly identical candidate set (27 to 28 accounts), differing only in how they partition the internal structure of the largest component.

Based on Table 5, internal cohesion across groups was calculated based on the safeguards specified in the Methods. Across the five two-account groups, pairwise similarities were tightly aligned between 0.854 and 0.991. Meanwhile, Group 2 the largest with 18 accounts recorded an average intra-group similarity of 0.715 (ranging down to 0.303), where 29 of 153 pair combinations cleared the threshold. This pattern validates a formation driven by transitive chaining rather than uniform pairwise closeness. Furthermore, complete-linkage decomposition revealed that Group 2 features eight tightly knit cores (one triplet and seven pairs meeting the threshold criteria) connected through bridging accounts. Thus, Group 2 is best understood as a collection of stylistically related sub-clusters instead of a single homogeneous entity.

Table 5. Identified Sockpuppet Groups

Group	Accounts	Highest Similarity	Status
Group 1	A1, B1	0.9085	Unreliable (thin evidence)
Group 2	A2, B2, C2, et al. (18 accounts)	0.9963	Campaign-template candidate
Group 3	A3, B3	0.8517	Unreliable (thin evidence)
Group 4	A4, B4	0.9511	Benign relay (false positive)
Group 5	A5, B5	0.9909	Confirmed coordinated distribution
Group 6	A6, B6	0.9297	False positive (syndication)

A systematic audit of all six groups against benign explanations, conducted in response to the topical-copy failure mode, produced the following taxonomy. Two groups carry campaign-template signatures: Group 2, whose members predominantly post short pro-revision slogans sharing an identical hashtag triplet across accounts, with several members posting duplicate texts in consecutive minutes, and Group 5, discussed below. Two groups are attributable to benign content syndication: Group 6, comprising two verified accounts of the same national media organization publishing identical editorial content, and Group 4, comprising two regional menfess-style relay accounts that broadcast a near-identical submitted message, a relay mechanism analogous to syndication. The remaining two groups (Groups 1 and 3) consist of colloquial, apparently organic accounts with very low tweet volumes (four and two tweets respectively, including single-tweet accounts below any reliable stylometric footing). Their groupings are treated as unreliable and excluded from coordination interpretation. This audit demonstrates that semantic similarity arising from shared sources is a systematic failure mode of the semantic-stylometric stage rather than an isolated exception, affecting at least two of the six groups, and that the mandatory elimination of legitimate explanations (criterion (d) of the minimum-evidence requirements) is an essential rather than optional component of the framework.

Group 5 (two accounts, hereafter Accounts A5 and B5) constitutes the strongest finding, supported by five independent evidence sources: BERT confidence scores of 0.9965 and 0.9964, a cosine similarity of 0.9909, graph edge existence, temporal co-posting within 13 seconds, and manual validation confirming that the accounts' tweets are verbatim identical with matching timelines. This evidence establishes coordinated content distribution through an automated or scheduled mechanism. It does not, by itself, establish common ownership or deceptive intent, which cannot be determined from platform data alone. The finding is therefore reported as high-confidence coordinated distribution consistent with, but not definitive proof of, inauthentic coordination.

3.4. Social Graph Analysis Results

The constructed graph contains 793 nodes and 737 weighted directed edges aggregated from the 1516 raw interactions, with an undirected density of 0.0023 as show in Fig. 2. The graph comprises 84 connected components, dominated by two large components (308 and 263 nodes) alongside numerous small star-shaped components; 681 nodes have a degree of one. Consequently, betweenness centrality is zero for the large majority of nodes, with 108 nodes exhibiting nonzero values (maximum 0.109); the highest-betweenness nodes correspond to organic conversation hubs rather than to any of the 28 candidate accounts, supporting the absence of broker accounts within the candidate network.

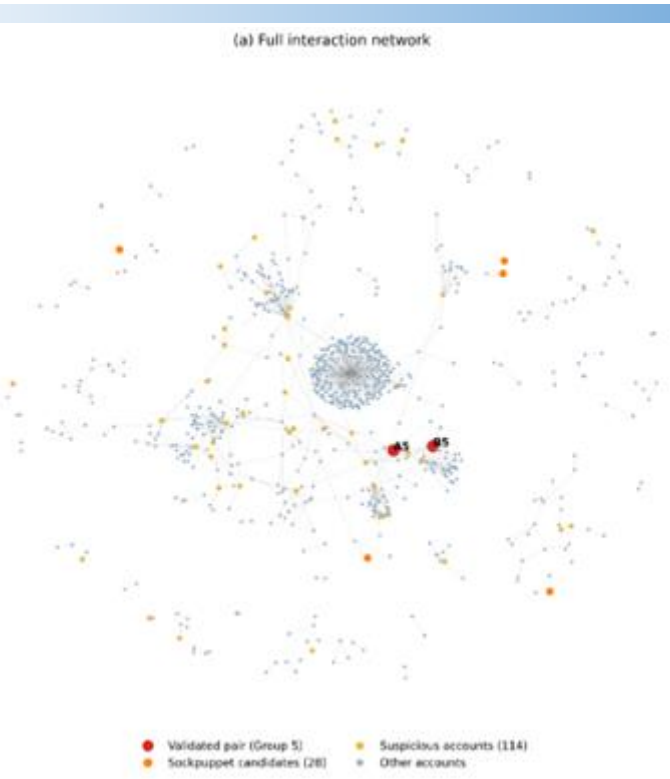

Fig. 2. Full Network Node, colors: red = validated pair (A5, B5); orange = 28 sockpuppet candidates; gold = 114 suspicious accounts; gray = other accounts

Community detection using the Louvain algorithm with a fixed random seed yielded 93 communities with a modularity score of 0.8277 as show on Fig. 3. This high modularity primarily reflects the sparsity and fragmentation of the interaction graph, in which many small, internally connected components are naturally well separated; it is a property of the graph structure under the chosen construction and is not, in itself, evidence of coordinated manipulation.

Fig. 3. Suspicious Subnetwork

As show on Fig. 4, zero coordinated clusters were found among the 28 candidate accounts under the two-of-three criteria defined in the Methods. The sensitivity analysis confirmed the robustness of this outcome, varying the bot-ratio threshold from 50% to 70% in 5% increments produced no change in the classification, as no community satisfied two or more criteria at any threshold value. The highest suspicious-account ratio observed across all communities with at least three members was 50% (a single four-account community), while communities of ten or more members exhibited ratios of at most 31%. Rather than clustering together, the candidate accounts were dispersed across multiple communities. Combined with the campaign-template evidence above, this dispersal is interpreted as consistent with implicit coordination that avoids direct interaction. However, dispersal alone is also compatible with independent actors posting similar content and this interpretation is applied only to the groups that carry independent corroborating evidence.

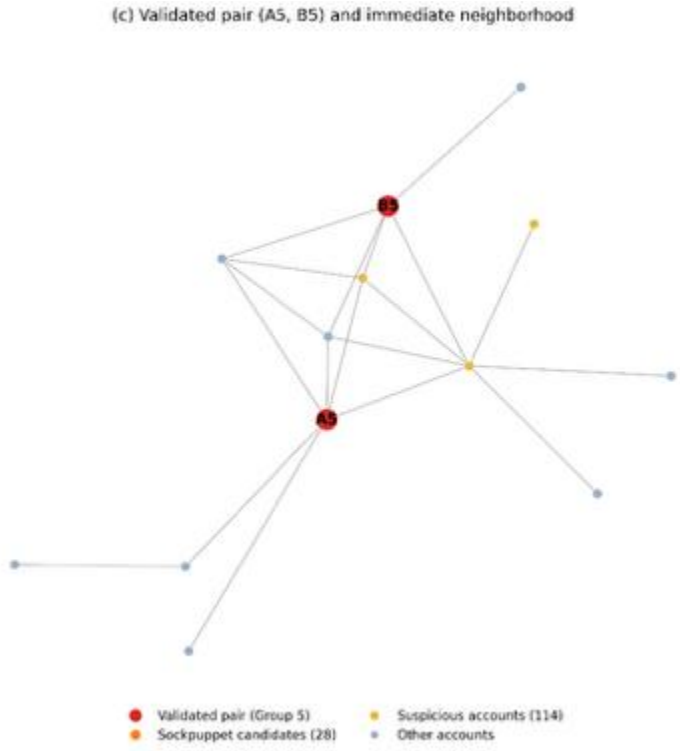

Fig. 4. Validated Pair

3.5. OSINT Investigation Results

Username lookup across TikTok, YouTube, and GitHub returned zero matching accounts, while Instagram and LinkedIn blocked automated queries. Consistent with the protocol stated in the Methods, these null results are reported as coverage limitations and are not interpreted as evidence of evasion.

Temporal analysis detected three synchronization slots in which more than one candidate account posted within the same minute. The permutation-based null model (10,000 iterations, preserving per-account tweet volumes and drawing timestamps from the empirical corpus distribution) shows that this count is not statistically unusual. 12.3 ± 2.8 slots would be expected by chance in this bursty six-day corpus. The raw slot count therefore carries no evidentiary weight. The evidentiary value lies at the pair level. All three co-posting events occurred exclusively between accounts of the same stylometric group (two within Group 2 and one within Group 5), whereas only 41.8% of candidate pairs are intra-group, and the Group 5 pair posted near-verbatim identical campaign text 13 seconds apart. Temporal co-posting is accordingly treated as a corroborating signal that aligns selectively with the stylometric grouping, not as standalone evidence of scheduled automation.

Deeper manual inspection of the Group 5 pair reveals that the full body text of both accounts' tweets is verbatim identical with zero lexical variation, deploying the same narrative framing and hashtags, with the only difference being hashtag placement, a superficial formatting artifact consistent with

template rendering rather than independent authorship as show on Fig. 5. Combined with same-date posting and the 13-second co-posting interval, this pair constitutes high-confidence evidence of coordinated automated content distribution within the candidate network, with the ownership and intent caveats stated above.

Fig. 5. Comparison of the validated pair (Accounts A5 and B5)

3.6. Framework Discussion

In the absence of target-domain ground truth, no end-to-end accuracy claim is made for the framework, and no quantitative superiority claim over single-method baselines is possible. What the results do support is a complementarity argument grounded in observed failure cases within this dataset: graph analysis alone detected no coordination, as the candidate accounts avoid direct interaction; semantic similarity alone produced systematic false positives, namely the syndication and relay groups; the BERT classifier alone flagged genuine passive accounts due to domain mismatch; and the raw temporal slot count alone was statistically unremarkable. Each single method failed in a documented way that a complementary stage corrected. The framework's practical contribution is therefore measured in investigative terms: it reduced the analyst workload from 720 accounts to 114 gated candidates, to 28 grouped candidates, and finally to a small number of pairs requiring manual inspection, while surfacing and eliminating false positives along the way.

The perfect benchmark scores warrant critical dissection. Feature-level analysis revealed two benchmark artifacts. First, the retweet_ratio feature perfectly separated the two classes, with a value of 0.0 for all genuine accounts versus 1.0 for all social spambots and zero variance in both classes. Inspection of the raw benchmark files shows that the underlying retweet field is fully populated in both classes. The separation therefore arose during feature construction, from the interaction between the presence-based feature logic and asymmetric ingestion of the two source files, rather than from genuine behavioral differences. Second, 10% of genuine accounts yielded empty aggregated text, a condition never observed in the bot class. Both artifacts constitute label-correlated shortcuts that a high-capacity classifier readily exploits.

An ablation experiment was therefore conducted in which both artifact features were removed and all empty-text accounts were excluded. The retrained model nevertheless achieved perfect scores on the held-out test partition (95 accounts). This indicates that the ceiling performance reflects the intrinsic homogeneity of the social spambots 1 subset, whose templated political content and uniform behavioral patterns are trivially separable for modern transformer architectures, rather than model superiority, consistent with the near-perfect account-level results reported on the Cresci-2017 benchmark by prior studies.

Kudugunta and Ferrara report 99.81% accuracy with precision, recall, and F1 rounding to 1.00 using an AdaBoost classifier with synthetic minority oversampling on the combined Cresci-2017 subsets [30] and BERT-based detectors similarly report high scores on this benchmark [31]. Furthermore, retraining the same architecture on the same data without a fixed weight-initialization seed produced divergent out-of-domain predictions for the 720 case accounts (114 versus 50 flagged accounts), despite both instances scoring perfectly on the benchmark. This constitutes direct evidence that benchmark performance masks substantial out-of-domain variance. Accordingly, the score of 1.0000 must be read strictly as evidence that the decision boundary of Cresci-2017 has been fully learned, not as a claim of equivalent performance on the Indonesian case data. The BERT classifier is treated

strictly as a first-stage gate whose outputs must survive independent corroboration by the subsequent stages, as demonstrated by the false positives surfaced and corrected in the group audit.

The hybrid framework's components are thus complementary rather than redundant based on Table 6. The weakness of each component is compensated by the strengths of the others, and evidence from four independent perspectives (text content, language style, network structure, and temporal patterns) converging on the same conclusion provides a substantially higher level of forensic confidence than any single method alone, subject to the evidentiary limits stated above.

Table 6. Framework Component Contributio

Component	Input	Key Output	Limitation
Preprocessing	1,309 raw tweets	1,303 clean tweets, 21 features, 1,516 edges	—
BERT Classification	720 accounts	114 suspicious candidates (15.8%); threshold-stable (0.60–0.95)	Domain mismatch (Cresci-2017); out-of-domain variance across retrained instances
Stylometry	114 accounts	28 candidate accounts in 6 groups; cohesion and audit taxonomy	Systematic topical-copy false positives (syndication and relay groups)
Graph Analysis	28 accounts + edges	0 coordinated clusters (robust at 50–70% threshold); dispersal consistent with implicit coordination	Implicit coordination undetectable by structure alone; sparse graph limits network signals
OSINT	28 accounts	3 sync slots (pair-level signal only); 1 pair confirmed; 1 false positive eliminated; forensic report	Raw slot count not statistically significant; Instagram and LinkedIn inaccessible

The most significant finding is the implicit nature of the observed coordination pattern. The corroborated groups do not interact directly but are identified through semantic-stylometric similarity and selective temporal co-posting, consistent with the implicit coordination strategies described in the coordinated behavior literature [7], [15], [32], subject to the evidentiary limits stated above.

4. CONCLUSION

This research developed and preliminarily evaluated a conditional hybrid framework for detecting coordinated suspicious accounts on the X platform by integrating BERT-based classification, semantic-stylometric analysis, social graph analysis, and OSINT investigation, addressing the methodological gap in which single-method approaches fail to detect implicit coordination patterns. Applied to 1,309 tweets from 720 accounts related to the Indonesian Military Law (UU TNI) issue in April 2025, the framework produced findings at four distinct levels of certainty: (1) at the classification level, 114 accounts (15.8%) were flagged as suspicious candidates; (2) at the grouping level, 28 accounts were organized into 6 similarity-based candidate groups, of which a systematic audit attributed two to benign content syndication and judged two more too thinly evidenced for interpretation; (3) at the coordination-evidence level, campaign-template content and selective intra-group temporal co-posting were observed in the two remaining groups; and (4) at the attribution level, one account pair was manually confirmed to distribute verbatim identical content through an automated or scheduled mechanism, with common ownership and deceptive intent remaining undetermined.

Because ground-truth labels are unavailable for the target accounts, end-to-end detection accuracy could not be established, and no quantitative superiority over single-method baselines is claimed. The framework's contribution is instead demonstrated through documented complementarity, in

which each single method failed in an observable way that another stage corrected, and through the reduction of analyst workload from 720 accounts to a small number of pairs requiring manual inspection. The most significant finding is the implicit nature of the corroborated coordination pattern. The accounts involved deliberately avoid direct interactions, rendering them undetectable by graph-based analysis alone, a pattern consistent with modern CIB strategies. This finding was reachable only through the combination of semantic-stylometric similarity, pair-level temporal analysis, and manual validation.

Several limitations bound these conclusions. The classification model was trained on Cresci-2017, which differs from the Indonesian 2025 target domain in language, period, and bot sophistication. 84.4% of target accounts contributed only a single tweet, limiting the reliability of individual stylometric fingerprints. Semantic similarity is conflated with shared topics and sources, producing systematic false positives that necessitate the elimination of legitimate explanations. The interaction graph is sparse and fragmented, constraining what network analysis can reveal. OSINT coverage was incomplete because two of the five platforms blocked automated queries. Manual validation was performed by a single annotator, and the publication of findings about real accounts carries ethical risks, addressed here through full anonymization and the accompanying ethics statement.

Future research should prioritize the construction of a manually labeled Indonesian-language benchmark to enable direct target-domain evaluation and blinded expert validation in which multiple independent annotators assess candidate accounts without knowledge of the system's outputs, enabling measurement of inter-annotator agreement and end-to-end precision. Further directions include domain adaptation of the classification stage, expanding OSINT coverage to platforms that restrict automated queries, and extending the temporal analysis to longer observation windows. Future work should also evaluate the scalability of the multi-stage architecture on significantly larger datasets, as the present evaluation relies on a relatively compact corpus of 1,309 tweets from 720 accounts. In particular, the pairwise stylometric comparison grows quadratically with the number of suspicious accounts, and although the conditional gating design already limits the volume of accounts entering the more computationally expensive stages, its effectiveness at the scale of nationwide campaigns involving tens of thousands of accounts remains to be validated, potentially requiring approximate nearest-neighbor search or blocking strategies to keep the comparison stage tractable. The proposed framework also holds potential for adaptation to other social media platforms and to emerging forms of AI-generated coordinated inauthentic behavior.

Ethics Statement

This study analyzes exclusively public tweets collected from the X platform. Because the classifications produced by the framework are candidate designations derived from a model trained on an out-of-domain benchmark, and because the group audit demonstrated that benign explanations account for a portion of the candidate groups, all account identities in this article are anonymized (accounts are referred to by group identifiers such as A5 and B5), and no account is labeled as fake or inauthentic as an established fact. Anonymization follows established guidance on the ethical handling of social media identities [33], [34] and is applied to prevent reputational harm to potentially misclassified genuine users. The de-anonymized mapping is retained by the authors and can be made available to the journal or to legitimate investigative authorities upon justified request. The authors assess that the public interest in methodological transparency for detecting coordinated manipulation outweighs the residual risks under these safeguards.

REFERENCES

[1] S. Gurajala, J. S. White, B. Hudson, B. R. Voter, and J. N. Matthews, “Profile characteristics of fake Twitter accounts,” *Big Data & Society*, vol. 3, no. 2, p. 2053951716674236, Dec. 2016, doi: 10.1177/2053951716674236.

[2] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, “The rise of social bots,” *Commun. ACM*, vol. 59, no. 7, pp. 96–104, Jun. 2016, doi: 10.1145/2818717.

[3] O. Varol, E. Ferrara, C. Davis, F. Menczer, and A. Flammini, "Online Human-Bot Interactions: Detection, Estimation, and Characterization," *ICWSM*, vol. 11, no. 1, pp. 280–289, May 2017, doi: 10.1609/icwsm.v11i1.14871.

[4] Meta Transparency Center, "Inauthentic Behavior | Transparency Center." Accessed: Jul. 18, 2026. [Online]. Available: https://transparency.meta.com/policies/community-standards/inauthentic-behavior

[5] F. Giglietto, N. Righetti, L. Rossi, and G. Marino, "It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections," *Information, communication & society : ICS ; an international journal for the Information Age*, vol. 23, no. 6, pp. 867–891, 2020, doi: 10.1080/1369118X.2020.1739732.

[6] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3980–3990. doi: 10.18653/v1/D19-1410.

[7] L. Mannocci, M. Mazza, A. Monreale, M. Tesconi, and S. Cresci, "Detection and Characterization of Coordinated Online Behavior: A Survey," 2024, *arXiv*. doi: 10.48550/ARXIV.2408.01257.

[8] A. Homsi, J. Al Nemri, N. Naimat, H. Abdul Kareem, M. Al-Fayoumi, and M. Abu Snober, "Detecting Twitter Fake Accounts using Machine Learning and Data Reduction Techniques:," in *Proceedings of the 10th International Conference on Data Science, Technology and Applications*, Online Streaming, --- Select a Country ---: SCITEPRESS - Science and Technology Publications, 2021, pp. 88–95. doi: 10.5220/0010604300880095.

[9] B. Bharti, N. S. Gill, and P. Gulia, "Exploring machine learning techniques for fake profile detection in online social networks," *IJECE*, vol. 13, no. 3, p. 2962, Jun. 2023, doi: 10.11591/ijece.v13i3.pp2962-2971.

[10] M. Mohammadrezaei, M. E. Shiri, and A. M. Rahmani, "Identifying Fake Accounts on Social Networks Based on Graph Analysis and Classification Algorithms," *Security and Communication Networks*, vol. 2018, pp. 1–8, Aug. 2018, doi: 10.1155/2018/5923156.

[11] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race," in *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, Perth, Australia: ACM Press, 2017, pp. 963–972. doi: 10.1145/3041021.3055135.

[12] S. R. Sahoo and B. B. Gupta, "Hybrid approach for detection of malicious profiles in twitter," *Computers & Electrical Engineering*, vol. 76, pp. 65–81, Jun. 2019, doi: 10.1016/j.compeleceng.2019.03.003.

[13] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, and M. Tesconi, "RTbust: Exploiting Temporal Patterns for Botnet Detection on Twitter," in *Proceedings of the 10th ACM Conference on Web Science*, Boston Massachusetts USA: ACM, Jun. 2019, pp. 183–192. doi: 10.1145/3292522.3326015.

[14] L. Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, and M. Tesconi, "Coordinated Behavior on Social Media in 2019 UK General Election," *ICWSM*, vol. 15, pp. 443–454, May 2021, doi: 10.1609/icwsm.v15i1.18074.

[15] F. Cinus, M. Minici, L. Luceri, and E. Ferrara, "Exposing Cross-Platform Coordinated Inauthentic Activity in the Run-Up to the 2024 U.S. Election," in *Proceedings of the ACM on Web Conference 2025*, Sydney NSW Australia: ACM, Apr. 2025, pp. 541–559. doi: 10.1145/3696410.3714698.

[16] D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer, "Uncovering Coordinated Networks on Social Media: Methods and Case Studies," *ICWSM*, vol. 15, pp. 455–466, May 2021, doi: 10.1609/icwsm.v15i1.18075.

[17] T. Solorio, R. Hasan, and M. Mizan, "A Case Study of Sockpuppet Detection in Wikipedia," in *Proc. Workshop on Language Analysis in Social Media (LASM)*, 2013, pp. 59–68.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," presented at the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[19] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, vol. 2008, no. 10, p. P10008, Oct. 2008, doi: 10.1088/1742-5468/2008/10/P10008.

[20] A. Alharbi, H. Dong, X. Yi, Z. Tari, and I. Khalil, "Social Media Identity Deception Detection: A Survey," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–35, Apr. 2022, doi: 10.1145/3446372.

[21] H. Satria, *tweet-harvest: A Twitter crawler*. (2025). GitHub repository. [GitHub repository]. Available: https://github.com/helmisatria/tweet-harvest

[22] J. Echeverría, E. De Cristofaro, N. Kourtellis, I. Leontiadis, G. Stringhini, and S. Zhou, "LOBO: Evaluation of Generalization Deficiencies in Twitter Bot Classifiers," in *Proceedings of the 34th Annual Computer Security Applications Conference*, San Juan PR USA: ACM, Dec. 2018, pp. 137–146. doi: 10.1145/3274694.3274738.

[23] C. Hays, Z. Schutzman, M. Raghavan, E. Walk, and P. Zimmer, "Simplistic Collection and Labeling Practices Limit the Utility of Benchmark Datasets for Twitter Bot Detection," in *Proceedings of the ACM Web Conference 2023*, Austin TX USA: ACM, Apr. 2023, pp. 3660–3669. doi: 10.1145/3543507.3583214.

[24] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[25] T. Pires, E. Schlinger, and D. Garrette, "How Multilingual is Multilingual BERT?," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, 2019, pp. 4996–5001. doi: 10.18653/v1/P19-1493.

[26] E. Stamatatos, "A survey of modern authorship attribution methods," *J. Am. Soc. Inf. Sci.*, vol. 60, no. 3, pp. 538–556, Mar. 2009, doi: 10.1002/asi.21001.

[27] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring Network Structure, Dynamics, and Function using NetworkX," presented at the Python in Science Conference, Pasadena, California, Jun. 2008, pp. 11–15. doi: 10.25080/TCWV9851.

[28] Techniques, I. I. E. C., "Iso/iec 27037: 2012 information technology security techniques guidelines for identification collection acquisition and preservation of digital evidence," ISO/IEC-The standard was published in October. Accessed: Jul. 19, 2026. [Online]. Available: https://iecstandardstore.com/

[29] K. Kent, S. Chevalier, T. Grance, and H. Dang, "Guide to integrating forensic techniques into incident response," National Institute of Standards and Technology, Gaithersburg, MD, NIST SP 800-86, 2006. doi: 10.6028/NIST.SP.800-86.

[30] S. Kudugunta and E. Ferrara, "Deep neural networks for bot detection," *Information Sciences*, vol. 467, pp. 312–322, Oct. 2018, doi: 10.1016/j.ins.2018.08.019.

[31] D. Dukic, D. Keca, and D. Stipic, "Are You Human? Detecting Bots on Twitter Using BERT," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, sydney, Australia: IEEE, Oct. 2020, pp. 631–636. doi: 10.1109/DSAA49011.2020.00089.

[32] L. Cima, L. Mannocci, M. Avvenuti, M. Tesconi, and S. Cresci, "Coordinated Behavior in Information Operations on Twitter," *IEEE Access*, vol. 12, pp. 61568–61585, 2024, doi: 10.1109/ACCESS.2024.3393482.

[33] A. S. Franzke, A. Bechmann, C. M. Ess, and M. Zimmer, "Internet Research: Ethical Guidelines 3.0," AoIR (The International Association of Internet Researchers), Report, 2020.

[34] C. Fiesler and N. Proferes, "'Participant' Perceptions of Twitter Research Ethics," *Social Media + Society*, vol. 4, no. 1, p. 2056305118763366, Jan. 2018, doi: 10.1177/2056305118763366.

AUTHOR BIOGRAPHY

RIKO IMAN DECAMARTA was born in Banjarbaru, South Borneo, Indonesia in 1994. He received the B.S. degree in nuclear engineering from the Sekolah Tinggi Teknologi Nuklir, Yogyakarta, in 2016. He has been pursuing the M.S. degree in Digital Forensics at Universitas Islam Indonesia, Yogyakarta, since 2022. His research interests include digital forensics, cybersecurity, and data science.

Yudi Prayudi is a lecturer in the Informatics Department at the Islamic University of Indonesia and currently serves as the Director of the Center for Digital Forensics Studies (CDFS). He earned a Doctoral degree in computer science in 2020 from the Department of Computer Science and Electronics, Gadjah Mada University, Yogyakarta. Previously, he obtained a Master of Informatics degree from the Sepuluh Nopember Institute of Technology, Surabaya, in 2002. and a Bachelor of Computer degree from Gadjah Mada University, Indonesia in 1994. His research interests are in digital forensics, digital evidence, cyberlaw, steganography, watermarking, and computer security. Some of the research topics he carries out include frameworks for digital forensics, chain of custody for digital evidence, blockchains for digital evidence, steganography and watermarking, and various forensic areas such as mobile forensics, cloud forensics, acquisitions, and multimedia forensics.
He is also active as an associate editor of the International Journal of Information Security and Privacy (IJISP), a reviewer board of the International Journal of Digital Crime and Forensics (IJDCF), the International Journal of Electronic Security and Digital Forensics (IJESDF), and the International Journal of Advanced Computer Science and Applications (IJACSA). He has also served as an expert witness in various legal cases involving computer crime. He is also currently listed as a national assessor for the National Accreditation Board for the Digital Forensics Testing Lab. He is also registered as a member of the Indonesian Digital Forensic Association and the High Technology Crime Investigation Association (HTCIA).