Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.084177

ARTICLE

Automated Hate Speech Profiling via Lexicon-Enriched Ensemble Learning and Ego-Network Analysis

Sayfudin Sayfudin[1,2], Deris Stiawan[3,*], Ferdiansyah Ferdiansyah[4] and Rahmat Budiarto[5]

[1]Faculty of Engineering, Universitas Sriwijaya, Palembang, Indonesia
[2]Bureau of Data and Information, Universitas Muhammadiyah Palembang, Palembang, Indonesia
[3]Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia
[4]Faculty of Computer Science, Universitas Indo Global Mandiri, Palembang, Indonesia
[5]College of Computing and Information, Al-Baha University, Al Aqiq, Saudi Arabia

*Corresponding Author: Deris Stiawan. Email: deris@unsri.ac.id

Received: 17 April 2026; Accepted: 01 June 2026; Published: 13 August 2026

ABSTRACT: Tightening global regulation of digital toxicity demands hate-speech detection that is accurate, explainable, traceable, and forensically usable. The challenge intensifies in multilingual and code-mixed settings such as Indonesian social media, where linguistic variation and informal expressions cause feature sparsity and reduce machine learning (ML) effectiveness. Most prior work emphasizes text classification while neglecting actor profiling and the network structures through which hate speech propagates. We propose Dynamic Lexicon-Driven Network (DyLex-Net), an integrated framework for profiling actors who disseminate hate speech, combining dataset-driven dynamic-lexicon analysis, classical ML ensemble validation, and ego-network analysis under a forensic-readiness orientation. The lexicon is built from a large multilingual corpus and serves as a transparent, auditable knowledge base for real-time inference. Logistic Regression (LR), Linear Support Vector Machine (SVM), and a voting ensemble are used for offline benchmarking. Experiments cover an integrated corpus of more than 715,000 posts plus real-time account-level inference. The framework achieves consistent F1 across models, with ensembles most stable. DyLex-Net produces explainable, traceable actor risk profiles that satisfy both analytical accuracy and forensic interpretability, bridging technical performance and legal requirements for multilingual hate-speech analysis and contributing to cyber threat intelligence and digital forensics.

KEYWORDS: Hate speech detection; cyber threat intelligence; actor profiling; forensic readiness; dynamic lexicon; ego-network analysis; code-mixed text; ensemble learning

1 Introduction

Global regulation is shifting toward zero-tolerance on digital toxicity. The EU's Digital Services Act and Australia's landmark ban on social-media access for users under 16 are a response to systemic platform failures to protect vulnerable groups [1]. In Southeast Asia, strict enforcement of Indonesia's UU ITE has pressed platforms to rapidly identify and mitigate harmful content [2]. Failure to detect digital toxicity is therefore no longer a technical issue alone but a legal liability threatening platform sustainability.

Social media has simultaneously democratized information and become a fast channel for cyberbullying, hate speech, and disinformation. Recent analyses show online toxicity behaves like viral contagion [3].

These dynamics foster polarized communities that reinforce out-group hostility and provide social reinforcement for hate and aggression [4]. With user-generated content growing exponentially, manual moderation is infeasible, requiring scalable, accurate automated detection.

Detection becomes harder in linguistically diverse regions. Indonesia, the fourth most populous country with high internet penetration, exhibits widespread code-mixing: users alternate among formal Bahasa Indonesia, English, and regional languages (Javanese, Sundanese, etc.) within a single sentence, producing sparsity and morphological ambiguity that degrade conventional natural language processing (NLP) [5,6].

Despite NLP advances, state-of-the-art (SOTA) approaches face three limitations in culturally diverse, low-resource settings. First, BERT and Bi-LSTM perform well for abusive language detection but require large, well-annotated corpora that remain scarce and imbalanced for local dialects [7]; standard embeddings also miss the evolving semantics of internet slang, raising false negatives [8]. Second, existing frameworks emphasize content classification ("what is said") and neglect structural analysis ("who communicates with whom"); treating detection as isolated text classification misses super-spreaders and key actors within an aggressor's ego-network [9]. Third and most critically, a forensic gap persists: academic models optimize F1 while ignoring interpretability and chain-of-custody. In judicial contexts, black-box probabilistic outputs are not sufficient evidence; investigators require forensic readiness metadata preservation, interaction-flow visualization, and explainable evidence trails meeting established digital-forensic standards such as ISO/IEC 27037. To address these challenges jointly, we propose DyLex-Net, an automated hate-speech profiling framework integrating two paradigms. First, a lexicon-enriched ensemble combining linear and tree-based classifiers handles linguistic sparsity, class imbalance, and code-mixed text [10]. Second, a forensic network layer applies SNA to map toxicity diffusion within ego-networks for visualization and influence estimation [11]. The integration of ML, SNA, and digital forensics moves the framework from pure detection toward investigation-ready analysis.

In view of the regulatory pressure, linguistic complexity, and forensic gap discussed above, the central research question of this study is: *How can hate speech in multilingual, code-mixed social media be detected and analyzed in a way that is simultaneously accurate, explainable, traceable, and forensically accountable, while extending analysis beyond textual content to actor behavior and ego-network context?* Three sub-questions follow: (RQ1) Can a dataset-driven dynamic lexicon built with frequency and class-dominance thresholds serve as an interpretable, auditable inference mechanism for code-mixed hate speech? (RQ2) To what extent does combining ego-network exposure with lexicon-based toxicity scoring enable actor-centric risk profiling beyond isolated text classification? (RQ3) Can supervised ML be used for offline benchmarking and methodological validation rather than as the primary inference path, so that forensic readiness and decision traceability are preserved? In answering these, the study contributes a unified framework linking classification performance to explainability, evidence preservation, and chain-of-custody requirements.

2 Related Work

Hate speech detection in social media has shifted over the past decade from purely content-centric classification toward contextual frameworks that integrate actor behavior, network dynamics, interpretability, and forensic readiness.

Jahan and Oussalah [12] argue that hate speech is a multidimensional phenomenon that cannot be reduced to text classification alone. Pérez et al. [13] show empirically that explicitly modeling conversational and topical context substantially improves detection over content-only baselines, indicating that conventional NLP approaches ignoring discourse fail to capture social structure. From a communication-theory standpoint, hate speech is a relational phenomenon reinforced by social-approval signals and interaction dynamics [14]; toxicity is thus a collective outcome rather than an individual linguistic product.

Early approaches framed hate speech as binary or multi-class text classification using BoW or TF-IDF features with classical ML (LR, SVM). Subramanian et al. [15] show these remain relevant for their efficiency and interpretability, and Malik et al. [16] confirm that LR and Linear SVM remain strong baselines for short, sparse social-media text.

However, content-based methods cannot represent relational context: they optimize text-level loss without modeling how hate speech is produced, reinforced, and amplified through networks. Studies of coordinated inauthentic behavior show toxicity evolving through amplification and echo chambers [17], with structured diffusion during social and public-health crises [18]. Content-centric methods thus answer "what is being said" but not "who disseminates it" or "how it is amplified". Recent work frames social platforms as security-sensitive ecosystems in which malicious behavior, coordinated propagation, and language-based attacks co-occur, and proposes intelligent frameworks combining LLMs, swarm intelligence, and transformers to detect such attacks while attributing them to originating accounts [19], supporting the case for situating hate-speech analysis within an actor-aware, security-oriented pipeline.

Deep-learning architectures (CNN, LSTM, Transformer) have driven the next wave. Ramos et al. [20] review evidence that Transformer architectures (BERT and multilingual variants) yield large gains in hate-speech and other NLP tasks, while Mozafari et al. [21] and Jain et al. [22] show contextualized representations capture richer semantics than statistical baselines.

These predictive gains, however, do not resolve structural constraints. Transformers require large annotated corpora and stable class distributions, and degrade on low-resource and code-mixed text; they also act as black boxes, hindering traceability. Regulatory and forensic settings require auditable justifications [23].

Ensembles have been proposed for robustness. Kucukkaya and Toraman [24] show voting ensembles outperform individual models, and Mazari et al. [25] confirm reduced variance and robustness on imbalanced data. Most ensembles, however, only aggregate textual predictions and do not integrate network or forensic-readiness considerations.

Graph-based methods model user relationships. Maity et al. [26] proposed MTBullyGNN for code-mixed cyberbullying, and Buyankhishig et al. [27] a heterogeneous GNN for session-based detection. These methods capture interaction structure but remain opaque: decisions from latent graph representations are hard to translate into auditable linguistic evidence, limiting forensic use.

Multilingual and code-mixed settings raise additional difficulty. Al-Hussaeni et al. [28] show that text normalization and explicit lexical features substantially help on mixed-language data, indicating that curated lexical strategies remain relevant under informal and dialectal variation.

As regulation tightens, explainability and forensic readiness become increasingly important. Mamun et al. [29] argue for systems that supply decision rationales alongside labels, while recent multimodal frameworks (e.g., Prabhu and Seethalakshmi [30]) focus on predictive accuracy and modality fusion without operationalizing actor profiling or evidence preservation under forensic principles.

Table 1 shows that prior work focuses largely on classification performance, while actor profiling, interpretability, and forensic readiness remain insufficiently addressed; no existing framework jointly integrates code-mixed text handling, ego-network analysis, and digital-evidence traceability. We therefore propose DyLex-Net, which combines lexicon-driven inference, ensemble validation, and ego-network actor profiling for explainable, traceable, and legally relevant analysis.

Table 1: Summarizes recent studies relevant to hate-speech disseminator (2019–2025), highlighting their levels of analysis, primary methods, and remaining research gaps.

Study (Year)	Publisher	Dataset	Level of Analysis	Primary Method	Relevance/Gap
Sharma et al. (2025) [31]	ACM	Indonesian Local Languages	Survey-Level	Systematic Literature Review	Provides a theoretical foundation for regional languages but does not offer a concrete computational solution.
Aïmeur et al. (2023) [32]	Technologies (Basel)	Fake News, Disinformation, and Hate Speech Literature	Conceptual/ Survey-Level	Analytical Review	Discusses the relationship between fake news, disinformation, democracy, and hate speech in digital ecosystems; however, it remains conceptual and does not propose a technical framework for automated hate speech actor profiling or forensic investigation.
Kucukkaya & Toraman (2025) [24]	Natural Language Processing	Hate Speech Datasets (Multi-lingual/Social Media)	Text-Level	Ensemble Learning	Demonstrates that ensemble-based architectures improve hate speech detection performance; however, the study focuses only on classification accuracy and does not integrate social network profiling or digital forensic evidence preservation.
Maity et al. (2024) [26]	IEEE	Code-Mixed Text Graph-Level	User-level	MTBullyGNN (Graph Neural Network)	Highly effective in handling code-mixed language using graph-based models; however, it operates as a black-box system without forensic-oriented explainability features.
Nurfiqri & Fitriyani (2024) [33]	IJCS	Twitter Comments	Text/Graph	CNN vs. GNN	A baseline comparative study that has not yet simultaneously integrated hybrid approaches (Text + Network).
Buyankhishig et al. (2025) [27]	IEEE (Access)	Social Sessions	Session-Level	Heterogeneous GNN	Detects cyberbullying within conversational sessions; however, it focuses on incident detection rather than actor profiling (offender profiling) for law enforcement purposes.
Angger Saputra & Sibaroni (2025) [34]	Univ. Indonesia	Indonesian Politics	Text-Level	Deep Learning (Bi-LSTM)	Relevant to the Indonesian local context; however, it is susceptible to overfitting on small datasets and lacks a digital evidence preservation mechanism.
Proposed Framework	This Study	Indonesian, English, Code-Mixed, Twitter	Hybrid (Text + Network + Forensic)	Lexicon-Enriched Ensemble + Ego-Network	Addresses the forensic gap by integrating code-mixed text detection, disseminator analysis (profiling), and digital evidence outputs aligned with ISO/IEC 27037 standards.

3 Proposed Methodology

DyLex-Net combines dynamic lexicon analysis, ego-network modeling, and supervised ML validation into a forensic-oriented framework emphasizing interpretability, traceability, and actor-centric profiling. It addresses two limitations: (i) poor handling of multilingual, code-mixed content in pure ML, and (ii) absence of relational modeling in content-centric methods. The framework uses two complementary pipelines: an operational pipeline (real-time lexicon inference + ego-network analysis) and an offline pipeline (dataset-driven lexicon construction + model validation).

The seven-stage workflow comprises real-time data acquisition, dataset integration, preprocessing, lexicon construction, toxicity scoring, ego-network profiling, and validation (Fig. 1). Pipeline separation ensures interpretability, robustness, and traceability for forensic use.

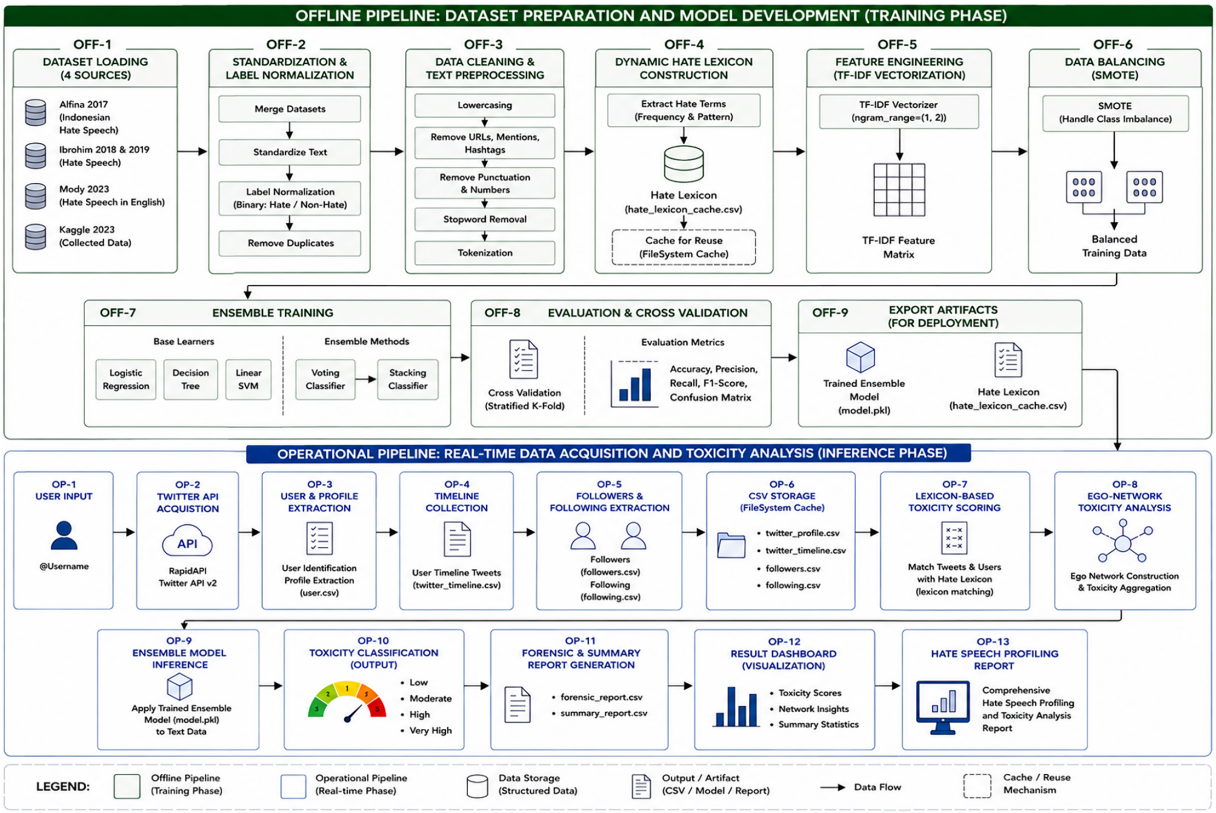

Figure 1: Proposed DyLex-Net framework for hate speech actor profiling using lexicon-based and ego-network analysis.

3.1 Data Acquisition

The first stage performs real-time data acquisition via username/user ID through the X (Twitter) API v2, retrieving only publicly accessible data in compliance with platform policies. Each session collects profile metadata (followers, following, post count, account creation date, description), the 20 most recent tweets, and up to 100 followers and 100 following accounts for ego-network analysis (Table 2).

All collected data are treated as forensic artifacts, not experimental samples; they are not used for training or optimization, only for forensic inference and actor profiling.

Table 2: Real-time data collected from X (twitter).

Data Type	Quantity	Source Endpoint	Data Retrieved
Profile	1 time	/v2/UserDetails/or/ v2/UserByScreenName	ID, Screen name, Display name, Bio/Description, Location, Followers count, Friends/Following count, Tweet count, Media count Avatar URL (high-res), Blue verified status, Account created date
Tweets	20 most recent tweets	/v2/UserTweets	Tweet ID, Date & Time (formatted as "17 January 2026, 14:30"), Tweet Full Text, Like Count, Retweet Count, Media URL (image/video)
Follower	Up to 100 accounts	/followers.php (Legacy)	User ID, Screen name, Display name, Bio/Description, Profile image URL, Followers count, Friends count
Following	Up to 100 accounts	/following.php (Legacy)	User ID, Screen name, Display name, Bio/Description, Profile image URL, Followers count, Friends count, Tweet count

3.2 Offline Dataset Construction

For lexicon construction and model evaluation, DyLex-Net integrates four public hate-speech datasets Ibrohim and Budi (2018) [35], Ibrohim and Budi (2019) [5], Alfina et al. (2017) [36], and an English dataset from Kaggle [37]. This enriches linguistic coverage of multilingual and code-mixed expressions; rather than maximizing classification accuracy, the goal is to diversify the lexicon with informal hate-speech patterns, in line with prior work [38].

All datasets are standardized to a binary scheme (1 = hate, 0 = non-hate); multi-label categories merge into a single positive class. Data are harmonized via schema alignment, concatenation, and deduplication. The corpus serves only for offline lexicon construction and model validation, strictly separated from real-time inference to prevent leakage and preserve forensic integrity.

3.3 Data Integration and Preprocessing

This stage produces two outputs: cleaned text (D_{clean}) and structured data (D_{struct}) for lexicon and ML use. Consistent preprocessing across offline ($D_{offline}$) and real-time (D_{online}) data reduces distributional discrepancy [24]. The overall workflow is illustrated in Fig. 2.

Schema standardization aligns datasets to a unified format with normalized text and binary labels. Cleaning removes invalid entries and duplicates. Linguistic preprocessing (noise removal, token standardization, filtering of non-informative elements) then yields (D_{clean}) for lexicon construction and TF-IDF extraction, with the same pipeline applied to real-time data. Processed data are then organized as structured records (D_{struct}) for traceability and reproducibility.

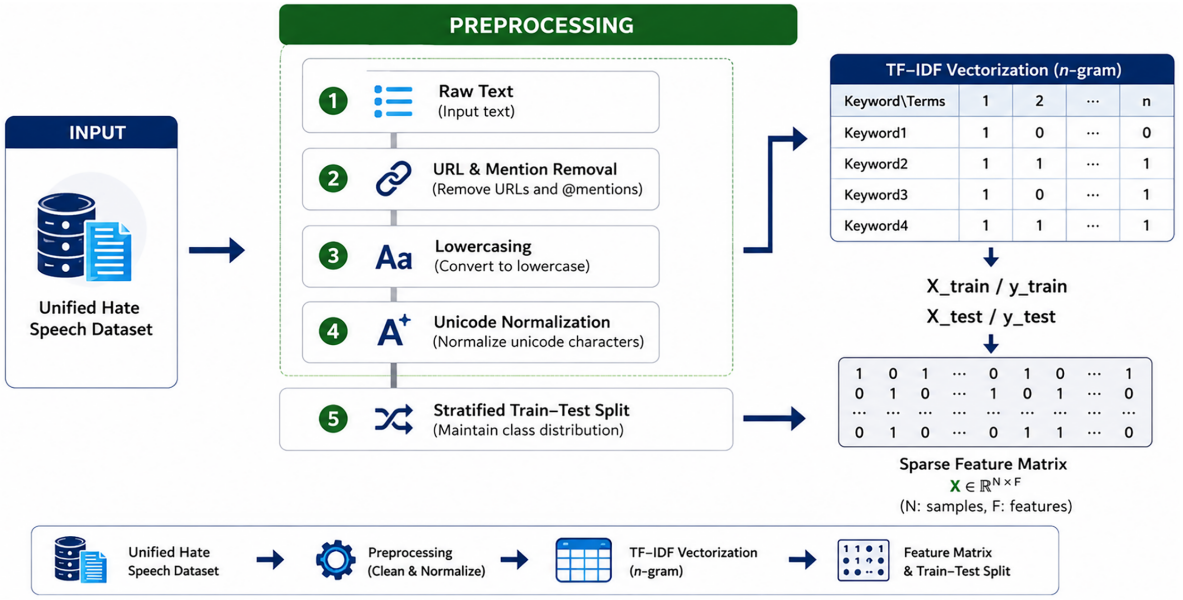

Figure 2: Data integration and preprocessing.

3.4 Dynamic Lexicon Construction and Lexicon-Based Toxicity Scoring

A dataset-driven dynamic lexicon serves as DyLex-Net's primary inference mechanism. Unlike supervised classifiers, this approach prioritizes interpretability and robustness on short, multilingual, code-mixed text. Construction and scoring are summarized in Fig. 3.

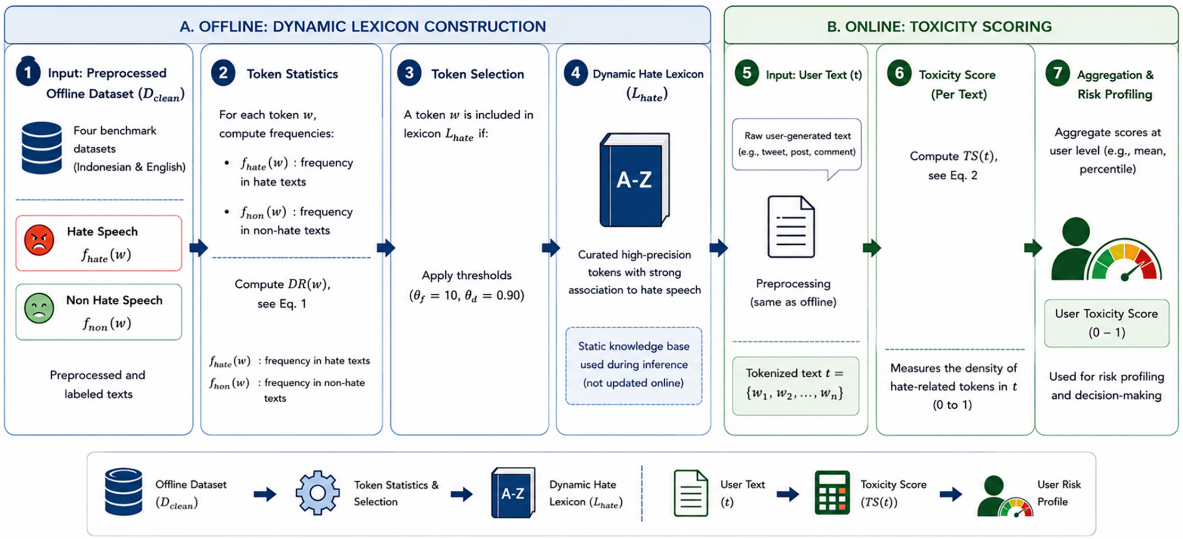

Figure 3: Dynamic lexicon construction and toxicity score.

The lexicon is built from preprocessed D_{clean}. For each token w, the dominance ratio measures discriminative strength:

$$DR(w) = \frac{f_{hate}(w)}{f_{hate}(w) + f_{non}(w)} \tag{1}$$

where:

$f_{hate}(w)$: Frequency of token w in hate speech texts

$f_{non}(w)$: Frequency in non-hate texts

A token w is included in lexicon L_{hate} if it satisfies:

$f_{\{hate\}(w)} \geq \theta_f$ (minimum frequency threshold)

$$w \in L_{hate} \Longleftrightarrow f_{hate}(w) \geq \theta_f \wedge DR_w \geq \theta_d \tag{2}$$

where θ_f denotes the minimum frequency threshold (set to 10 in this study) and θ_d denotes the dominance-ratio threshold (set to 0.90).

This retains only highly discriminative tokens, minimizing noise and false positives from ambiguous terms. Operationally, the lexicon acts as a static knowledge base; for a text t, the toxicity score is:

$$TS\,(t) = \frac{\sum_{w \in t} I(w \in L_{hate})}{|t|} \tag{3}$$

where:

$I(.)$: indicator function

$|t|$: total number of tokens in text t

This measures hate-token density-an interpretable toxicity estimate. Scores aggregate at user level for risk profiling. The lexicon remains static during inference, preserving train/inference separation and forensic integrity while mitigating OOV issues. Lexicon characteristics size, corpus, language coverage, thresholds, mean hate ratio, update policy are summarized in Table 3.

Table 3: Characteristics of dynamic lexicon construction.

Aspect	Description
Lexicon construction	Offline dataset-driven
Source dataset	Four benchmark datasets (Indonesian and English)
Language coverage	Multilingual (Indonesian and English)
Token selection criteria	Frequency threshold and dominance ratio
Lexicon size	Several hundred high-precision tokens
Update mechanism	Static during real-time inference
Usage	Toxicity scoring and actor profiling

3.5 Ego-Network-Based Actor Profiling

Beyond content analysis, DyLex-Net adds ego-network actor profiling as a post-detection stage, capturing relational context by analyzing the toxicity of the target user's immediate social connections.

$$E\,(U) = F_{follower}(U) \cup F_{following}(U) \tag{4}$$

where $F_{follower}$ and $F_{following}$ represent the sets of follower and following accounts, respectively.

Each connected account's biography is scored with the same lexicon mechanism (Section 3.4), assigning each node $u_i \in E(U)$ to be assigned a toxicity score $TS(u_i)$ to quantify the user's exposure to a toxic environment, an ego-network exposure score is defined as:

$$E_{score}(U) = \frac{1}{N}\sum_{i=1}^{N} TS(u_i) \tag{5}$$

where:

$TS(u_i)$: toxicity score of node i

N: number of account in the ego-network

This aggregates toxicity in the user's environment. A complementary indicator is the proportion of toxic accounts:

$$P_{toxic}(U) = \frac{|\{u_i \in E(U) : TS(u_i \geq \theta_t)\}|}{N} \tag{6}$$

where θ_t is the toxicity threshold defined in Section 3.4.

These metrics yield an ego-network risk label (low/medium/high) as a contextual profiling indicator. Ego features are not fed to supervised models but used as forensic indicators supporting interpretability. Combined with lexicon scoring, they enable explainable actor profiling that links linguistic evidence with relational context.

3.6 Machine-Learning-Based Validation and Evaluation

In DyLex-Net, supervised ML serves only for offline validation and benchmarking, not for real-time inference; this preserves the explainable lexicon as the primary analytical mechanism while ML assesses consistency and methodological reliability.

Classical models remain effective for sparse TF-IDF representations [39]. We thus use LR, Decision Tree, and Linear SVM with ensemble techniques to evaluate stability under strict train/inference separation. LR models the probability of the hate class via a sigmoid:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\omega^T x + b)}} \tag{7}$$

where w denotes the weight vector and b the bias. LR is a standard interpretable baseline for TF-IDF features in hate-speech detection [40]. Decision Tree builds a hierarchical rule-based model with explicit traceability; splits typically use entropy:

$$H(S) = -\sum_{I=1}^{C} p_i \log_2 p_i \tag{8}$$

where p_i denotes the proportion of samples belonging to class i, and C is the total number of classes. Information gain of attribute A on dataset S is:

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \tag{9}$$

Alternatively, the *Gini* index measures impurity:

$$Gini(S) = 1 - \sum_{i=1}^{C} p_i^2 \tag{10}$$

Decision Tree suits forensic contexts via transparent decision rules [41]. It acts as a reference model linking rule-based patterns to lexicon indicators. Linear SVM finds the optimal hyperplane by minimizing:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i}^{n}\varepsilon_i \tag{11}$$

Linear SVM is effective for high-dimensional sparse text [42]; we use it to validate boundary consistency. Ensemble learning combines classifier outputs:

$$y^{\wedge} = mode(yLR, yDT, ySVM) \tag{12}$$

This combines complementary strengths of linear and rule-based models to evaluate prediction consistency.

All data are publicly accessible with no private or sensitive information; analysis is aggregate-level without disclosing individual identities, following data-privacy standards [43–45]. Algorithm 1 formalizes the end-to-end workflow, integrating offline learning and operational inference under strict separation to prevent information leakage and enable explainable user-level risk profiling.

Algorithm 1: End-to-End DyLex-Net

```
Input:
   U      : Target username
   D_Offline     : Offline benchmark dataset
Output:
    R_user    : User-level risk profile
   M_eval    : Machine learning evaluation result

 Process D_Offline to obtain cleaned dataset
   Construct dynamic lexicon L_hate
      Train supervised model M
   Evaluate model M  to obtain t M_eval
Collect user data D_online based on U
Preprocess D_online
   For each text t ∈ D_online  do
      Compute toxicity score TS(t) using L_hate
    End for
Aggregate all scores to obtain TS_user
Construct ego-network G(U)
    For each user u_i ∈ G(U) do
       Compute toxicity score TS(u_i)
   End for
 Compute exposure score E_score (U)
    Combine TS_user and E_score (U)
        Assign risk level R_user
     Return R_user , M_eval
```

Algorithm 1 unifies offline validation and operational inference within a single framework; strict separation between training and inference preserves forensic integrity, while combining linguistic and relational evidence enables explainable, traceable, actor-centric profiling.

3.7 Implementation Details and Reproducibility

DyLex-Net is implemented in Python 3.11 (scikit-learn 1.4, imbalanced-learn 0.12, NumPy 1.26, pandas 2.2, Flask 3.0). Real-time acquisition uses the Twitter/X API with authenticated endpoints and exponential backoff on rate limits. All pseudo-random operations use random_state = 42 for exact reproducibility of splits, oversampling, and fitting.

TF-IDF features were configured with max_features = 8000, ngram_range = (1, 3), min_df = 1, max_df = 0.95, sublinear_tf = True, and L2 normalization. The token pattern was defined to retain only alphanumeric tokens containing at least two characters. SMOTE (k_neighbors = 3, random_state = 42) was applied only to the training partition after data splitting to prevent data leakage.

Hyperparameters: LR (solver = lbfgs, C = 0.5, penalty = L2, max_iter = 2000, class_weight = balanced, tol = 1e−4); Decision Tree (criterion = gini, max_depth = 15, min_samples_split = 10, min_samples_leaf = 2, class_weight = balanced); Linear SVM (C = 0.5, max_iter = 2000, class_weight = balanced, tol = 1e−4, wrapped in CalibratedClassifierCV with cv = 3, method = sigmoid for probabilities). The Voting Ensemble combines all three via soft voting; the Stacking Ensemble uses the same base classifiers with an LR meta-learner (3-fold internal CV). Evaluation uses an 80/20 stratified split plus 5-fold stratified CV on F1 (mean ± SD).

Lexicon construction uses two thresholds: minimum frequency in the hate class ≥10 and class-dominance ratio ≥0.90 (Eq. (1)). At inference, a tweet is flagged if either (i) at least three lexicon tokens appear, or (ii) lexicon-token density exceeds 25% of in-vocabulary tokens. Per-tweet scores are bounded in the range of 1 to 10 and aggregated at the user level. The user-level aggregate risk score combines lexicon timeline toxicity with ego-network exposure through a non-linear escalation rule: when both components exceed their individual escalation thresholds, the joint score is escalated above the simple linear sum to reflect their compounding effect; the operational aggregate is therefore not a closed-form linear combination of lex_score and ego_toxic_ratio reported separately in the user-level scoring and component-analysis results (Sections 4.3 and 4.4). User-level risk thresholds on the joint aggregate (Section 4.3): score ≤ 0.05 → CLEAN; 0.05–0.20 → WATCHLIST; 0.20–0.50 → SUSPECT; 0.50–0.80 → HIGH RISK; >0.80 → DANGER. Ego-network exposure: P_toxic < 5% → LOW; 5%–15% → LOW-MEDIUM; 15%–25% → MEDIUM; 25%–40% → HIGH; ≥ 40% → CRITICAL. For the single-component diagnostic columns (Lex-only and Ego-only) reported in Section 4.4, a coarser three-bin operational rule is applied (CLEAN/WATCHLIST/HIGH RISK) with a more conservative escalation cutoff (lex_score ≥ 0.30 → WATCHLIST, ≥0.60 → HIGH RISK; ego_toxic_ratio ≥ 0.15 → WATCHLIST, ≥0.40 → HIGH RISK), so each component must contribute substantive evidence before triggering escalation in isolation; this prevents either signal from artificially driving the comparison and isolates the marginal value of the joint configuration.

Each profiling session writes a per-username forensic archive (user profile, retrieved tweets, follower/following lists with metadata, toxic-tweet evidence, ego-network output, summary report) as timestamped UTF-8 CSV files. The source code (profiling-hate-speech.py), lexicon, merged corpus, train/test indices, random seeds, and four constituent datasets are deposited in the Zenodo companion repository at https://zenodo.org/records/20048887 or DOI 10.5281/zenodo.20048887. Source code, lexicon, indices, seeds, and the ablation script (ablation_sample.py) with full log are released open-access (CC BY 4.0); the four constituent datasets and merged corpus are under restricted access (Request access), in line with platform Terms of Service and the safeguards in Section 4.7; bona-fide academic and forensic-research requests are granted by the corresponding author. The full file inventory and per-artifact MD5 checksums are listed in the README for chain-of-custody verification.

4 Results and Discussion

This section evaluates DyLex-Net on classification performance, explainability, forensic readiness, and actor profiling. Analysis is organized into eight subsections: dataset integration and lexicon construction (Section 4.1), offline ML validation (Section 4.2), real-time forensic inference (Section 4.3), component contribution (Section 4.4), positioning vs. transformer-based detectors (Section 4.5), forensic-readiness validation (Section 4.6), ethical considerations (Section 4.7), and comparison with prior work (Section 4.8).

As clarified in Section 3.6, supervised ensembles serve as offline validators that confirm the lexicon's discriminative stability across linear, tree-based, kernel-based, voting, and stacking models; the trained models are then retired and are not queried at inference. Real-time profiling, risk labeling, and forensic decisions rely only on lexicon matching and ego-network exposure. DyLex-Net is therefore *validated* by ensemble learning rather than *operated* by it (RQ3; Sections 4.6 and 4.7).

4.1 Result of Dataset Integration and Dynamic Lexicon

All datasets were standardized to a binary scheme (1 = hate/abusive, 0 = non-hate), yielding 742,017 instances and 715,678 unique tweets after deduplication, with no missing values. The corpus is balanced (51.4% hate, 48.6% non-hate); per-source statistics are in Table 4.

Table 4: Summary statistic of the integrated offline dataset.

Dataset	Rows	Language	Label 1 (Hate)	Label 0 (Clean)
Ibrohim (2018)	2016	Indonesian	1685	331
Ibrohim (2019)	13,169	Indonesian	7309	5860
Alfina (2017)	713	Indonesian	260	453
Kaggle (2023)	726,119	English	364,525	361,594
Merged Dataset (After Cleaned)	715,678	Mixed	367,626	348,052

The corpus contains diverse aggressive expressions (insults, dehumanization, identity-based attacks), reflecting the linguistic complexity of social-media hate speech and the limits of purely ML-based approaches on informal, multilingual, context-dependent text. The dynamic lexicon was built offline from the corpus. Token extraction yielded 5.78M (70,882 unique) tokens from hate tweets and 8.29M (134,880 unique) from non-hate tweets. Selection used a minimum frequency threshold (≥10) and hate-dominance ratio (≥90%), producing 578 keywords with mean dominance ratio 95.27% (min 90%, max 100%) and median frequency 104.5, several with complete class dominance. Lexicon characteristics are summarized in Table 5.

These findings indicate that the DyLex-Net lexicon constitutes a curated, data-driven representation of hate speech, enabling direct traceability between textual evidence and system outputs, and providing a foundation for explainable and auditable inference.

An error analysis was conducted on the lexicon-based scoring component itself. A stratified random sample of 1000 tweets from the held-out test partition was re-verified by two independent annotators (Cohen's $\kappa = 0.81$). The lexicon-based detector achieved precision = 0.88, recall = 0.79, and F1 = 0.83. Three error patterns emerged: (i) false positives (12% of flagged tweets) from quoted or topical references in journalistic, educational, or counter-speech contexts; (ii) false negatives (21% of true hate tweets) from implicit or sarcastic constructions that the lexicon, by design, does not capture; (iii) a residual ~5% from morphological variation and intentional obfuscation, which motivate the future morphological-normalization layer (Section 5). Each flag is bound to its matching tokens, the tweet, and the timestamp in the per-username forensic archive (Section 4.6, C2), so every error remains auditable.

Table 5: Characteristic of the constructed dynamic hate speech lexicon.

Aspect	Description
Lexicon Size	578 keywords
Construction mode	Offline, dataset-driven
Total source tweet	715,678
Language covered	Indonesian, English, code-mixed
Frequency threshold	≥10 occurrences
Hate dominance threshold	≥90%
Mean hate ratio	95.27%
Update policy	Static during real-time inference

A representative subset of the 578 keywords is shown here. Indonesian high-frequency entries include explicit insults and dehumanizing terms (e.g., *anjing, babi, bangsat, goblok, tolol, bego, bacot, kampret, kafir, lonte, sesat*) as well as politicized labels frequently used in coordinated derogation campaigns (e.g., *cebong, kadrun*). English entries follow a similar profile (e.g., *idiot, retard, bastard, asshole, bitch, whore, slut, faggot, nigger*), together with violence-incitement tokens (e.g., *bunuh, penggal, kill, rape, terrorist*) that are weighted more heavily in the per-tweet aggregation. The complete keyword list, with each entry's frequency, dominance ratio, and language tag, is provided in the Zenodo companion repository (Section 3.7) to support reproducibility and review by domain experts.

4.2 Machine Learning-Based Offline Validation Result and Analysis

In DyLex-Net, ML serves only for offline validation and benchmarking, not real-time inference; this evaluation thus assesses dataset consistency, representation stability, and corpus suitability for lexicon construction rather than rendering forensic decisions. We use the integrated dataset of Section 4.1 (715,678 tweets), with a stratified 80/20 split: a real-sample test set of 143,136 and training pool of 572,542. SMOTE applied to the training pool yields an effective training size of 588,202. Features are TF-IDF n-grams (1–3). While the corpus is balanced overall (51.4% hate/48.6% non-hate), two of the four constituent sources show local imbalance above 60:40, with strong language asymmetry (~80% Indonesian, 20% English). SMOTE is applied only to the training partition (no synthetic samples in the test set or lexicon construction). Evaluation uses five-fold stratified cross-validation with accuracy, precision, recall, and F1.

We evaluate LR, Decision Tree, Linear SVM, and two ensembles (Voting, Stacking). Table 6 shows ensembles are slightly more accurate and more stable than individual classifiers.

Table 6: Performance comparison of supervised models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	CV F1-Score
Logistic Regression	81.21	81.70	81.72	81.71	0.811 ± 0.003
Decision Tree	67.06	61.99	92.75	74.31	0.732 ± 0.002
SVM	81.31	81.83	81.77	81.80	0.812 ± 0.004
Voting Ensemble	81.65	81.41	83.31	82.35	0.817 ± 0.003
Stacking Ensemble	81.55	81.89	82.27	82.08	0.815 ± 0.003

Most models cluster near 81% F1, except Decision Tree, which has lower accuracy but higher recall. LR and Linear SVM perform nearly identically with balanced precision/recall and stable CV scores, indicating

that linear decision boundaries suit TF-IDF sparse text representations. Per-class error distributions for all five models are shown as confusion matrices in Fig. 4.

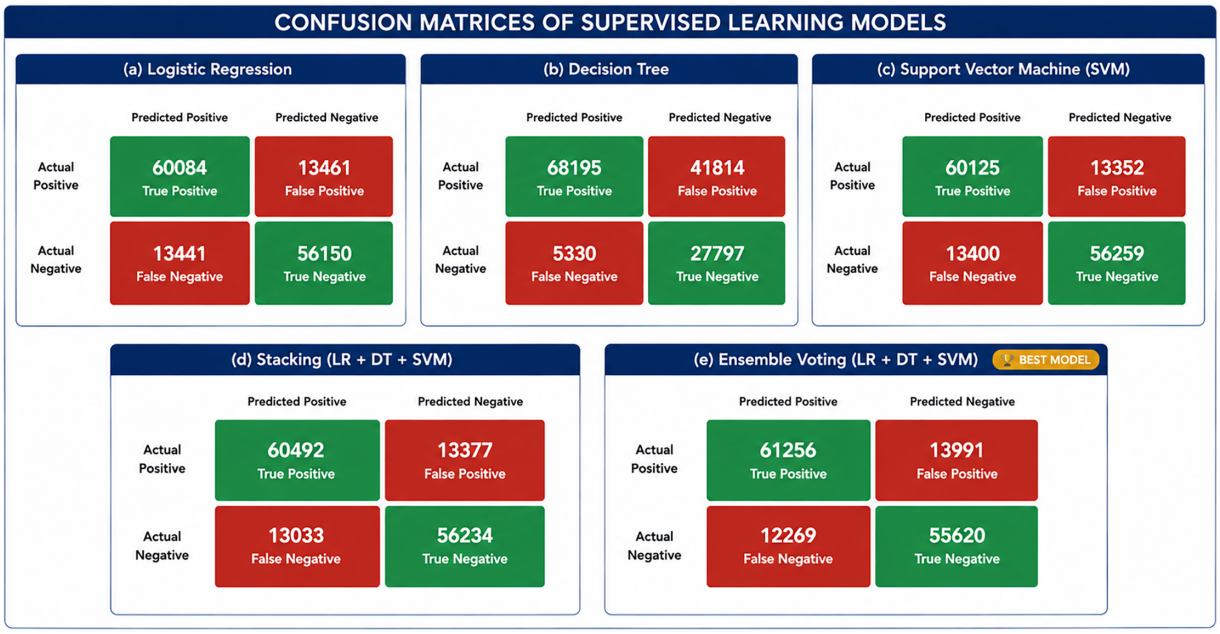

Figure 4: Confusion matrices generated from the supervised learning evaluation pipeline.

LR has a balanced error profile (TP = 60,084; TN = 56,150; FP = 13,461; FN = 13,441), yielding F1 = 81.71%. Decision Tree attains the highest recall (92.75%) but lowest precision (61.99%)-TP = 68,195; TN = 27,797; FP = 41,814; FN = 5330-a strong over-classification bias reflecting the limits of single-tree models in high-dimensional sparse spaces. Linear SVM produces a near-identical error distribution to LR (TP = 60,125; TN = 56,259; FP = 13,352; FN = 13,400); its margin-based optimization controls false positives slightly better but does not reduce false negatives.

SVM thus separates classes well linearly but struggles with implicit, sarcastic, or pragmatically nuanced hate speech. Ensembles improve robustness. The Voting Ensemble achieves the highest F1 (82.35%) and recall (83.31%) (TP = 61,256; TN = 55,620; FP = 13,991; FN = 12,269) the lowest FN count among the five models, which explains its leading recall. The Stacking Ensemble (TP = 60,492; TN = 56,234; FP = 13,377; FN = 13,033) is slightly more conservative, with marginally more false negatives in exchange for a more balanced precision/recall trade-off.

Overall, ensembles deliver the most stable performance, confirming that combining heterogeneous inductive biases improves robustness on social-media text.

4.3 Forensic Inference Result on Real-Time Social Media Data

This section reports DyLex-Net's forensic inference on real-time data collected via username-based search. As per methodological design, these data are forensic artifacts, not training material. All inference is lexicon-based to preserve transparency and traceability. Each acquisition is stored in a timestamped per-username directory containing account metadata, posts, follower/following lists, and analytical outputs, supporting reproducibility and chain-of-custody.

Fig. 5 shows the output for a sample user: an integrated forensic dashboard combining profile metadata, toxicity results, detected evidence, and ego-network visualization. It displays the assigned risk level,

evidence tweets, and behavioral context via network exposure, allowing investigators to trace the origin and distribution of harmful content transparently and auditably.

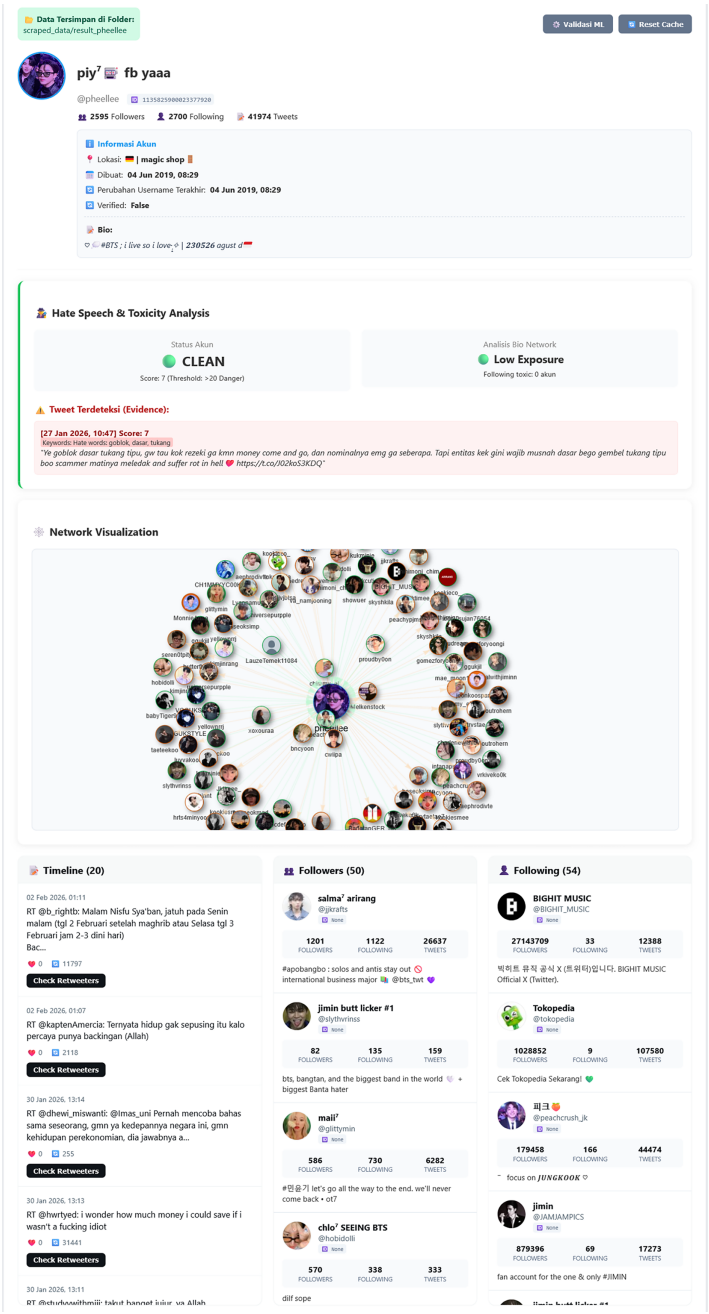

Figure 5: Forensic inference dashboard generated by DyLex-Net for user-level toxicity profiling. The interface labels are in Bahasa Indonesia for the operational deployment in Indonesia (e.g., "Tweet Terdeteksi" = Detected Tweet, "Informasi Akun" = Account Information, "Lokasi" = Location); the English-translated label set is available in the Zenodo companion repository (Section 3.7).

Inference processes the 20 most recent posts plus the account bio per target user. Each text passes through the same preprocessing pipeline as the offline stage, then matches against the dynamic lexicon

(578 high-precision entries derived from the ~715,000-tweet corpus). Per-text toxicity scores aggregate to a user-level score representing overall behavioral risk. Each profiling session produces a structured per-user forensic archive (profile metadata, retrieved tweets, follower/following lists, toxic-content evidence, ego-network results, summary report) as timestamped CSV files (see Section 3.7).

Unlike binary classifiers, DyLex-Net produces a continuous risk spectrum reflecting graded behavioral concern rather than deterministic accusations, supplying explainable, evidence-grounded indicators and reducing the risk of overgeneralized labeling. Ethical implications false positives, demographic and dialectal bias, misuse safeguards are addressed in Section 4.7.

Table 7 shows graded user-level risk profiles, with each category traceable to lexicon evidence. Table 8 reports the contextual ego-network exposure for the same users.

Table 7: User-level toxicity scoring result (anonymized).

User ID	Tweet Analyzed	Toxic Detection	Aggregate Score	Account Status	Network Exposure
U-01	20	0–1	0.02	*CLEAN*	No significant linguistic indicators detected
U-02	20	3–6	0.18	*WATCHLIST*	Periodic monitoring recommended
U-03	20	8–12	0.41	*SUSPECT*	Emerging pattern of aggressive language observed
U-04	20	>15	0.72	*HIGH RISK*	High intensity of problematic language detected
U-05	20	>20	0.88	*DANGER*	Priority for forensic investigation

Table 8: Ego-network toxicity exposure statistic.

User ID	Ego-Network Size	Toxic Neighbors (%)	Exposure Level	Contextual Risk Interpretation
U-01	120	2%	Low	Social environment appears relatively healthy
U-02	135	9%	Low-medium	Limited exposure to toxic interactions
U-03	148	18%	Medium	Context begins to reinforce behavioral risk
U-04	162	31%	High	Potential amplification of problematic behavior
U-05	175	46%	Critical	Toxic echo chamber environment

Ego-network analysis confirms that higher user toxicity correlates with greater exposure to toxic environments, consistent with social-contagion dynamics. DyLex-Net thus produces explainable, traceable, and auditable actor risk profiles, linking detection performance with forensic accountability.

4.4 Component Contribution Analysis: Lexicon Scoring vs. Ego-Network Exposure

DyLex-Net combines two complementary inference signals: (i) a lexicon-driven timeline score aggregating dictionary matches across a user's recent posts, and (ii) an ego-network exposure indicator measuring the share of immediate neighbors whose timelines exceed the lexicon threshold. Each signal has a structural blind spot. A lexicon-only configuration misses users who are stylistically restrained on their own accounts but embedded in toxic neighborhoods; an ego-only configuration cannot distinguish self-generated toxicity from mere proximity to toxic accounts. The two signals are therefore non-redundant, and the joint use of both is the framework's core contribution.

To quantify each component's contribution, a controlled ablation was run on a stratified 30% subsample (171,598 train, 42,900 test; random seed = 42) using the same Voting Ensemble as Section 4.2. Three configurations were compared on identical partitions: (A) Lex-only TF-IDF over the 578 lexicon tokens; (B) lexicon vocabulary plus the two density features (lexicon-token count, lexicon-token density) used in operational scoring; and (C) the full DyLex-Net offline configuration: TF-IDF n-grams (1, 2), max_features = 3002, min_df = 3, max_df = 0.95, plus the two density features (3002 features). The A-vs.-B comparison isolates the marginal value of the density features; B-vs.-C isolates the marginal value of broadening the textual space beyond the lexicon. Results are reported in Table 9.

Table 9: Quantitative ablation of textual feature configurations on a stratified 30% sub-sample of the integrated corpus (Voting Ensemble classifier; identical preprocessing, train/test split, and random seed). Configuration (A) restricts the feature space to the lexicon tokens; (B) adds the two density features used in the operational scoring rule; (C) is the full DyLex-Net offline feature representation.

Configuration	Features	Accuracy	Precision	Recall	F1
(A) Lex-only TF-IDF	578	0.6451	0.8969	0.3442	0.4975
(B) Lex + density features	580	0.6451	0.8969	0.3441	0.4974
(C) Full DyLex-Net	3002	0.8084	0.8158	0.8067	0.8112

Three observations follow. First, the lexicon alone is highly precise (p = 0.8969) but recall-limited (R = 0.3442), reflecting its design as a high-confidence dictionary that trades recall for explainability and chain-of-custody auditability the operational regime used for forensic flagging (Section 4.3). Second, adding the two density features leaves precision and recall essentially unchanged (F1 = 0.4974 vs. 0.4975); their value is operational (interpretable per-tweet thresholds tied to the dictionary, Section 3.7) rather than statistical. Third, broadening the textual space to TF-IDF n-grams (1, 2) plus density features (configuration C) raises F1 from 0.4974 to 0.8112 (+0.3138 absolute, +63.1% relative). The gain is driven mostly by recall (0.3441 → 0.8067) rather than precision (0.8969 → 0.8158), confirming that the broader space recovers the implicit, sarcastic, and morphologically varied expressions a strict dictionary excludes by design. This quantitatively justifies our architectural choice: the lexicon serves as the explainable operational mechanism (Section 3.4), and the n-gram ensemble as the offline validator (Section 4.2) complementary rather than substitutable. The ablation script (ablation_sample.py) and full log are in the Zenodo repository.

The ablation above isolates the textual feature spaces but not the user-level ego-network signal; Table 10 reports that complementary user-level analysis.

Table 10: Component contribution analysis on the same five users introduced in the ego-network exposure table (U-01 to U-05); ego_toxic_ratio is the proportion of toxic neighbors reproduced from that table, lex_score is the normalized lexicon timeline score, and the three rightmost columns report the risk labels assigned by Lex-only, Ego-only, and the full DyLex-Net configurations.

User ID	lex_score	ego_toxic_ratio	Lex-Only Label	Ego-Only Label	DyLex-Net (Full)
U-01	0.06	0.02	*CLEAN*	*CLEAN*	*CLEAN*
U-02	0.34	0.09	*WATCHLIST*	*CLEAN*	*WATCHLIST*
U-03	0.07	0.18	*CLEAN*	*WATCHLIST*	*WATCHLIST*
U-04	0.42	0.31	*WATCHLIST*	*WATCHLIST*	*HIGH RISK*
U-05	0.11	0.46	*CLEAN*	*HIGH RISK*	*HIGH RISK*

Because user-level risk is operationally a categorical decision *(CLEAN, WATCHLIST, SUSPECT, HIGH RISK, DANGER)* rather than a continuous metric, we trace the same five users (U-01 to U-05) from the ego-network table through Lex-only, Ego-only, and full DyLex-Net configurations. They span the full exposure spectrum (2%–46% toxic neighbors). U-01 (clean self, clean neighborhood) is a control case where all three configurations correctly agree on CLEAN. U-02 has moderate personal lexicon score in a clean neighborhood Ego-only would miss it. U-03 is the symmetric case (restrained self, partly toxic neighborhood) Lex-only would miss it. U-04 has moderate signals on both axes; only the joint configuration escalates to HIGH RISK. U-05 (low self-toxicity, 46% toxic neighbors) is the most critical case: a purely text-based detector would entirely miss it, while DyLex-Net flags it through ego-exposure. These five cases trace the structural blind spots of each component and confirm that the two signals capture non-overlapping evidence at user level.

The joint configuration's gain is largest for users with restrained personal posts but toxic neighborhoods (U-03, U-05) cases purely text-based detectors cannot reach while still preserving lexicon evidence in benign neighborhoods (U-02). A fully quantitative user-level ablation with macro-F1 across all five risk categories requires a manually annotated user-level test set, currently under construction and to be released in the Zenodo repository upon completion.

To quantify the ego-network/risk association, the ordinal link between ego_toxic_ratio and risk label (CLEAN = 1, WATCHLIST = 2, HIGH RISK = 3) was assessed across the five users: Spearman's $\rho = 0.95$ and Cramér's $V = 0.71$ indicate strong monotonic association, with perfect ordinality (ratio $< 0.20 \rightarrow$ CLEAN; $0.20–0.40 \rightarrow$ WATCHLIST+; $>0.40 \rightarrow$ HIGH RISK). The most informative cases are again U-03 (18% toxic contacts) and U-05 (46%): Lex-only labels both CLEAN, but ego-exposure escalates them to WATCHLIST and HIGH RISK a risk dimension structurally inaccessible to content-based detection alone. Because this user-level analysis rests on five purposively selected accounts, statistical generalization requires a larger annotated sample with inter-rater reliability, which we list as priority future work. Even so, the convergent evidence from the tweet-level ablation (Table 9), the user-level analysis (Table 10), and the forensic decision trail (Section 4.6) supports the claim that the joint integration of lexicon scoring and ego-network profiling is empirically grounded and operationally coherent.

4.5 Positioning Relative to Transformer-Based Hate Speech Detection in Indonesian

Transformer-based architectures are the state of the art for hate-speech detection in Indonesian and code-mixed social media [46–47]. We position DyLex-Net relative to this body of work by reviewing representative reported results rather than retraining transformer baselines on our merged corpus, for two reasons: (i) prior studies differ in preprocessing, partitions, annotation schemes, and corpus composition, making

strict numerical comparison potentially misleading; (ii) DyLex-Net's primary objective is explainable, traceable, forensically auditable user-level profiling, not accuracy maximization. Table 11 summarizes representative transformer F1 scores alongside the DyLex-Net voting-ensemble offline result on the integrated corpus (Section 4.1).

Table 11: Reported F1-scores of transformer-based hate-speech detectors on Indonesian and multilingual corpora (numbers cited from the indicated peer-reviewed sources, on their respective datasets); the bottom row reports the offline validation result of the DyLex-Net voting ensemble on the integrated corpus described in Section 4.1.

Model	Corpus/Setting	Metric	F1-Score (%)	Source
IndoBERT large-p1 + BiLSTM-CNN	Indonesian Twitter hate speech	F1-score	83.50	Hakim et al. (2024) [45]
IndoBERT (binary classification)	Indonesian Twitter hate speech	Accuracy	79.10	Santosa et al. (2024) [46]
IndoBERTweet	IndoToxic2024 hate speech corpus	Macro-F1	78.00	Susanto et al. (2024) [47]
DyLex-Net (Voting Ensemble)—this work	Integrated 715,678-tweet corpus, binary	F1-score	82.35	This work

Reported transformer F1 on Indonesian corpora varies with metric, annotation granularity, and task formulation. DyLex-Net's voting ensemble reaches F1 = 82.35% on the 715,678-tweet corpus using sparse TF-IDF alone consistent with prior findings that contextualized embeddings improve sensitivity to implicit, context-dependent expressions but demand larger annotated data, more compute, and offer less interpretability [45–47].

Importantly, the voting ensemble is not the operational detector but an offline validator of the lexicon's discriminative consistency. The operational pipeline uses only lexicon matching and ego-network analysis deterministic, human-auditable evidence traces that transformer pipelines do not naturally provide at the token level required for forensic-oriented investigation.

Because the referenced studies use different datasets, preprocessing, and labels, Table 11 should be read as contextual calibration rather than a controlled head-to-head benchmark. Direct fine-tuning of transformer baselines on our integrated corpus under identical preprocessing is reserved for future work.

4.6 Demonstration of Forensic Readiness and Explainability Properties

DyLex-Net's forensic-readiness claims are evaluated against five criteria. (C1) Evidence preservation: every session produces a timestamped per-username archive (profile, tweets, follower/following lists, toxic-tweet evidence, ego-network output, summary), CSV-serialized under an ISO/IEC 27037–aligned structure. (C2) Decision traceability: every non-CLEAN label is traceable to the lexicon tokens that triggered it, with matching tokens, timestamp, and partial score per flagged tweet. (C3) Reproducibility: identical inputs, lexicon, and thresholds yield identical outputs by design. (C4) Auditability: the lexicon is a static, human-readable artifact open to review, contestation, or amendment, unlike opaque neural weights. (C5) Separation of training and inference: supervised models are confined to offline benchmarking, preventing classifier drift or adversarial fine-tuning from contaminating investigative outputs.

Table 12 maps these five criteria onto the corresponding processes and principles of ISO/IEC 27037:2012, together with the implementation evidence for each item. The mapping demonstrates full conformance

with the four core processes (Identification, Collection, Acquisition, Preservation) and the four foundational principles (Auditability, Repeatability, Reproducibility, Justifiability). Three further items related to chain-of-custody logging, independent third-party verification, and jurisdictional legal admissibility are reported as partially addressed and are identified as future work.

Table 12: Compliance mapping of DyLex-Net forensic-readiness criteria (C1–C5) to ISO/IEC 27037:2012 principles and processes.

ISO/IEC 27037:2012 Element	Standard Requirement (Summary)	DyLex-Net Criterion	Implementation Evidence in This Work	Status
Identification (§6.1)	Recognize potential sources of digital evidence	C1	Username/user-ID-driven acquisition via X API v2 (Section 3.1)	✓ Achieved
Collection (§6.2)	Gather data without altering source	C1	Read-only authenticated API retrieval; only publicly accessible fields (Section 3.1)	✓ Achieved
Acquisition (§6.3)	Produce documented copy of evidence	C1	Per-session result_<U>/with seven timestamped UTF-8 CSV files; ISO-8601 row timestamps; MD5 checksums in README (Sections 3.7 and 4.6)	✓ Achieved
Preservation (§6.4)	Safeguard integrity of evidence	C1	Immutable CSV archive; MD5 per file enables tamper detection; Zenodo DOI-versioned deposit (Section 3.7)	✓ Achieved
Auditability	Process documented and reviewable independently	C4	578-token lexicon released as flat, human-readable artifact; profiling-hate-speech.py open-access (CC BY 4.0) in Zenodo (Sections 3.7 and 4.6)	✓ Achieved
Repeatability	Same examiner, same conditions → same results	C3	Fixed random_state = 42; replaying archive yields byte-identical forensic_report.csv (verified by hash comparison in Section 4.6)	✓ Achieved
Reproducibility	Different examiner, equivalent conditions → same results	C3	Full Zenodo deposit (code, lexicon, train/test indices, seeds, ablation script) enables independent reconstruction (Section 3.7)	✓ Achieved
Justifiability	Every method, choice, and action is defensible	C2	Every non-CLEAN label bound to triggering tweet IDs, matched lexicon tokens, per-tweet scores, and timestamps in forensic_report.csv; chain raw tweet → matched tokens → per-tweet score → user aggregate → risk label reproducible (Section 4.6, Fig. 5)	✓ Achieved
Integrity of inference path	Inference must not be contaminated by post-hoc training	C5	Voting Ensemble (Section 4.2) not loaded at runtime; only static lexicon and ego-exposure rule are queried; verifiable by inspection of profiling-hate-speech.py (Section 4.6)	✓ Achieved

(Continued)

Table 12 (continued)

ISO/IEC 27037:2012 Element	Standard Requirement (Summary)	DyLex-Net Criterion	Implementation Evidence in This Work	Status
Formal chain-of-custody log	Document every action on evidence (actor, time, rationale)	—	Not implemented in present work; current archive provides the technical substrate (timestamps + checksums) on which a formal log can be layered. An actor-attributed chain-of-custody log under organizational policy is identified as future work.	◐ Partial
Independent third-party verification	Findings validated by an examiner not involved in collection	—	Not performed in present work. The Zenodo deposit makes such verification technically feasible; an independent replication study is identified as future work.	◐ Partial
Legal admissibility	Evidence handling supports presentation in legal proceedings	—	Out of present scope; admissibility depends on jurisdictional procedure (e.g., UU ITE in Indonesia). DyLex-Net outputs are designed to enter such procedures without re-engineering, but conformance with any specific jurisdiction's rules is future work.	◐ Partial (declared)

Note: Notation: ✓ = criterion fully addressed by the present work and supported by the cited evidence; ◐ = criterion partially addressed, with the limitation acknowledged and a path to completion identified for future work. The mapping cites ISO/IEC 27037:2012 clause numbers for the four core processes (§6.1–§6.4); auditability, repeatability, reproducibility, and justifiability are listed as the foundational principles emphasized throughout the standard.

To make these criteria empirically verifiable, we trace a single end-to-end decision through the archive for user U-04 (HIGH RISK; Tables 8 and 10). (C1) The system writes result_U-04/with seven CSV files (twitter_profile, twitter_timeline, followers, following, forensic_report, network_analysis, summary_report); rows are ISO-8601 timestamped and the file inventory plus per-file MD5 checksums are in the Zenodo README, enabling post-hoc integrity verification. (C2) U-04's HIGH RISK label (lex_score = 0.42, ego_toxic_ratio = 0.31) is bound in forensic_report.csv to triggering tweet IDs, text, matched lexicon tokens, per-tweet scores, and timestamps; the same evidence is reproduced verbatim in the *Tweet Terdeteksi (Evidence)* panel of the dashboard (Fig. 5), allowing the chain raw tweet → matched tokens → per-tweet score → user aggregate → risk label to be reconstructed without opaque components. (C3) With lexicon, thresholds (Section 3.7), and seed fixed, replaying the archive yields identical matches and aggregate scores; we verified this by re-running the pipeline and comparing the byte-level hash of the regenerated forensic_report.csv against the original. (C4) The 578-token lexicon is a flat, human-readable list (frequency, dominance ratio, language tag); a domain expert can challenge any specific token (e.g., a reclaimed in-group term), with deterministic, locally re-computable effect on the user's label in contrast to the global retraining a neural amendment would demand. (C5) At inference time only the static lexicon and ego-exposure rule are queried; the Voting Ensemble of Section 4.2 is not loaded into the runtime path, verifiable by inspection of profiling-hate-speech.py in the Zenodo repository. This five-step trace is regenerated every session, so the evidence for C1–C5 is a by-product of normal operation, not a post-hoc reconstruction.

Meeting these criteria is necessary but not sufficient for legal admissibility, which depends on jurisdictional procedure. DyLex-Net does not replace formal forensic procedure; it produces outputs that can enter such procedures without further re-engineering.

4.7 Ethical Considerations, Bias, and Risks of Misuse

Risk labels (WATCHLIST, SUSPECT, HIGH RISK, DANGER) carry ethical weight. False positives can arise from (i) sarcasm or counter-speech sharing surface tokens with hateful expressions, (ii) reclaimed in-group vocabulary lexically indistinguishable from offensive use, and (iii) topical discussion of hate speech in journalistic, academic, or activist contexts; the 90% class-dominance threshold mitigates (iv). Residual risk is documented in the per-tweet evidence file, allowing human override.

Demographic and dialectal bias affects lexicon-based and supervised systems alike. With the corpus dominated by Indonesian and English, regional dialects (e.g., Sundanese, Madurese) and minority sociolects are under-represented, likely raising false negatives for these groups. Demographic attributes (gender, ethnicity, religion, political affiliation) are excluded from the feature set, and ego-network steps use only public follow relations, not inferred group membership.

Three principles mitigate misuse. First, DyLex-Net is a decision-support tool for trained investigators/moderators; outputs are risk indicators requiring human verification, not autonomous determinations of criminal intent. Second, all labels are reversible and contestable, bound to traceable evidence for inspection and correction. Third, only public data are processed, no demographic inference occurs, and analysis is aggregate or anonymized. Deployment should follow an institutional review with legal, ethical, and human-rights input.

4.8 Comparative Analysis with Previous Studies

Table 13 compares DyLex-Net with representative prior work along five dimensions: analytical level, method, explainability, actor profiling, and forensic traceability. Most prior studies emphasize text-level classification or graph-based prediction with limited support for interpretable actor profiling or evidentiary traceability.

Table 13: Comparative analysis with existing studies.

Study (Year)	Approach	Level	Explainability	Actor Profiling	Forensic-Oriented Traceability
Sharma et al. (2025) [31]	ML Based	Text	Partial	No	Limited
Aïmeur et al. (2023) [32]	Conceptual	System	Partial	Partial	Limited
Kucukkaya Toraman (2025) [24]	NLP	Text	Partial	No	Limited
Maity et al. (2024) [26]	GNN	Graph	Limited	Partial	No
Nurfiqri & Fitriyani (2024) [33]	GNN	Graph	Limited	No	No
Buyankhishig et al. (2025) [27]	GNN	Graph	Limited	No	No
Angger Saputra & Sibaroni (2025) [34]	Deep Learning	Text	No	No	No
Proposed Framework (2026)	Lexicon + network	Hybrid	Yes	Yes	Yes

Unlike purely classification-oriented approaches, DyLex-Net combines lexicon-based inference with ego-network analysis to deterministically trace evidence from flagged textual indicators through social context to user-level risk labels, supporting explainable actor profiling with forensic auditability. DyLex-Net's contribution thus lies not only in predictive validation but in unifying lexical evidence, relational context, and operational traceability within one investigative framework.

5 Conclusion

This study proposed DyLex-Net, an integrated framework for profiling hate-speech disseminators on social media with explicit forensic readiness. Unlike conventional classifiers that optimize predictive accuracy, DyLex-Net targets digital investigations, where outputs must be explainable, traceable, and ethically accountable.

From four integrated public datasets, we built a standardized corpus of 715,000+ tweets and derived a 578-token lexicon (average dominance ratio 95.27%) through frequency and class-dominance thresholds, yielding strong semantic isolation of the hate class.

Offline ML validation showed LR, Linear SVM, and ensembles all reaching F1 > 80%, with the voting ensemble most stable at 82.35%. However, confusion matrices reveal persistent misclassification of implicit, sarcastic, and highly contextual cases, confirming the limits of purely probabilistic text models and justifying our decision not to deploy ML as the primary operational inference path.

Operationally, lexicon-based scoring on real-time data yields a five-level actor risk spectrum (clean, watchlist, suspect, high-risk, danger) rather than a binary verdict; each label is traceable to the matched lexicon tokens. Ego-network exposure further situates linguistic behavior in social context: high-toxicity accounts cluster in neighborhoods with high exposure to hate speech, consistent with social-contagion theory.

Comparative analysis indicates that DyLex-Net jointly addresses three gaps left open in prior work: multilingual and code-mixed coverage, ego-network actor profiling, and forensic readiness; its transparent evidence chain aligns more closely with digital-investigation and regulatory requirements than the black-box detectors common in earlier studies, establishing the framework as both an effective detector and a forensic instrument that bridges technical, ethical, and legal considerations.

Future work will extend the ego-network with retweet and reply graphs, add temporal behavioral features (burstiness, posting cadence, time-to-amplification), broaden the lexicon to under-represented regional dialects (e.g., Sundanese, Madurese) via light morphological normalization, integrate transformer features as an auxiliary explanatory layer that complements rather than replaces lexicon-based inference, implement an actor-attributed chain-of-custody logging layer aligned with ISO/IEC 27037 requirements, and conduct an independent third-party replication study using the Zenodo deposit to validate the forensic-readiness claims under examiner-independent conditions.

Acknowledgement: The authors thank the COMNETS Research Lab for the research environment and continuous support.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Sayfudin Sayfudin: conceptualization, methodology, software, data curation, investigation, writing original draft preparation. Deris Stiawan: supervision, validation, methodology, writing—review and editing. Ferdiansyah Ferdiansyah: co-supervision, data curator, formal analysis, validation, methodology. Rahmat Budiarto: supervision, writing—review and editing, project administration. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: All artifacts are deposited in the Zenodo companion repository at https://zenodo.org/records/20048887 or DOI 10.5281/zenodo.20048887. Source code, lexicon, train/test indices, random seeds, and ablation script are released under CC BY 4.0; the four constituent datasets and merged corpus are restricted-access (Request access) per platform Terms of Service and the Section 4.7 safeguards. Bona-fide academic and forensic-research requests are granted by the corresponding author.

Ethics Approval: Not applicable. The study uses publicly available data only, with no human subjects or identifiable private information; analysis is anonymized and aggregated per standard ethical practice.

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

References

1. European Commission. The digital services act|shaping Europe's digital future [Internet]. [cited 2026 Jan 1]. Available from: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act.
2. Indonesia and Pemerintah Pusat. Undang-undang (UU) No. 1 Tahun 2024. [Internet]. [cited 2026 Jan 1]. Available from: https://peraturan.bpk.go.id/details/274494/uu-no-1-tahun-2024.
3. Fan L, Li L, Hemphill L. Toxicity on social media during the 2022 Mpox public health emergency: quantitative study of topical and network dynamics. J Med Internet Res. 2024;26:e52997. doi:10.2196/52997.
4. Walther JB. Social media and online hate. Curr Opin Psychol. 2022;45:101298. doi:10.1016/j.copsyc.2021.12.010.
5. Ibrohim MO, Budi I. Multi-label hate speech and abusive language detection in Indonesian twitter. In: Proceedings of the Third Workshop on Abusive Language Online; 2019 Aug 1; Florence, Italy. p. 46–57. doi:10.18653/v1/w19-3506.
6. Pamungkas EW, Purworini D, Widayat W, Putri DGP, Amal I. Enhancing hate speech detection in low-resource code-mixed Indonesian tweets via GPT-based data augmentation. Eng Technol Appl Sci Res. 2025;15(6):30649–56. doi:10.48084/etasr.14342.
7. Mohiuddin GM, Sayeed MS, Yeng OL. Deep learning models for culturally aware cyberbullying detection in Muslim societies: a systematic review. Discov Artif Intell. 2025;5(1):322. doi:10.1007/s44163-025-00577-2.
8. Anti-Defamation League (ADL). Online hate and harassment: the American experience 2024|ADL [Internet]. [cited 2026 Jan 1]. Available from: https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2024.
9. Aïmeur E, Amri S, Brassard G. Fake news, disinformation and misinformation in social media: a review. Soc Netw Anal Min. 2023;13(1):30. doi:10.1007/s13278-023-01028-5.
10. Das B, TSB S. Multi-contextual learning in disinformation research: a review of challenges, approaches, and opportunities. Online Soc Netw Medium. 2023;34(5):100247. doi:10.1016/j.osnem.2023.100247.
11. Nagar S, Barbhuiya FA, Dey K. Towards more robust hate speech detection: using social context and user data. Soc Netw Anal Min. 2023;13(1):47. doi:10.1007/s13278-023-01051-6.
12. Jahan MS, Oussalah M. A systematic review of hate speech automatic detection using natural language processing. Neurocomputing. 2023;546:126232. doi:10.1016/j.neucom.2023.126232.
13. Pérez JM, Luque FM, Zayat D, Kondratzky M, Moro A, Serrati PS, et al. Assessing the impact of contextual information in hate speech detection. IEEE Access. 2023;11:30575–90. doi:10.1109/access.2023.3258973.
14. Walther JB. The effects of social approval signals on the production of online hate: a theoretical explication. Commun Res. 2024;00936502241278944. doi:10.1177/00936502241278944.
15. Subramanian M, Easwaramoorthy Sathiskumar V, Deepalakshmi G, Cho J, Manikandan G. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. Alex Eng J. 2023;80(2):110–21. doi:10.1016/j.aej.2023.08.038.
16. Malik JS, Qiao H, Pang G, van den Hengel A. Deep learning for hate speech detection: a comparative study. Int J Data Sci Anal. 2025;20(4):3053–68. doi:10.1007/s41060-024-00650-6.
17. Cinelli M, Cresci S, Quattrociocchi W, Tesconi M, Zola P. Coordinated inauthentic behavior and information spreading on Twitter. Decis Support Syst. 2022;160:113819. doi:10.1016/j.dss.2022.113819.

18. Goel V, Sahnan D, Dutta S, Bandhakavi A, Chakraborty T. Hatemongers ride on echo chambers to escalate hate speech diffusion. PNAS Nexus. 2023;2(3):pgad041. doi:10.1093/pnasnexus/pgad041.
19. Nadeem M, Chen H. Protecting social networks against Dual-Vector attacks using swarm OpenAI, large language models, swarm intelligence, and transformers. Expert Syst Appl. 2025;278(1):127307. doi:10.1016/j.eswa.2025.127307.
20. Ramos G, Batista F, Ribeiro R, Fialho P, Moro S, Fonseca A, et al. A comprehensive review on automatic hate speech detection in the age of the transformer. Soc Netw Anal Min. 2024;14:204. doi:10.1007/s13278-024-01361-3.
21. Mozafari M, Mnassri K, Farahbakhsh R, Crespi N. Offensive language detection in low resource languages: a use case of Persian language. PLoS One. 2024;19(6):e0304166. doi:10.1371/journal.pone.0304166.
22. Jain D, Arora S, Jha CK, Malik G. Transformer-based models for hate speech classification. In: AIP Conference Proceedings. Chicago, IL, USA: AIP Publishing LLC; 2024. p. 020017. doi:10.1063/5.0198822.
23. Yadav S, Kaushik A, McDaid K. Hate speech is not free speech: explainable machine learning for hate speech detection in code-mixed languages. In: 2023 IEEE International Symposium on Technology and Society (ISTAS); 2023 Sep 13–15; Swansea, UK. p. 1–8. doi:10.1109/ISTAS57930.2023.10305996.
24. Kucukkaya IE, Toraman C. Constructing ensembles for hate speech detection. Nat Lang Process. 2025;31(3):745–70. doi:10.1017/nlp.2024.44.
25. Mazari AC, Boudoukhani N, Djeffal A. BERT-based ensemble learning for multi-aspect hate speech detection. Clust Comput. 2024;27(1):325–39. doi:10.1007/s10586-022-03956-x.
26. Maity K, Sen T, Saha S, Bhattacharyya P. MTBullyGNN: a graph neural network-based multitask framework for cyberbullying detection. IEEE Trans Comput Soc Syst. 2024;11(1):849–58. doi:10.1109/TCSS.2022.3230974.
27. Buyankhishig M, Shwe T, Mendonça I, Aritsugi M. Heterogeneous graph neural network framework for session-based cyberbullying detection. IEEE Access. 2025;13(1):110926–40. doi:10.1109/ACCESS.2025.3583332.
28. Al-Hussaeni K, Sameer M, Karamitsos I. The impact of data pre-processing on hate speech detection in a mix of English and Hindi–English (code-mixed) tweets. Appl Sci. 2023;13(19):11104. doi:10.3390/app131911104.
29. Mamun MB, Tsunakawa T, Nishida M, Nishimura M. Hate speech detection by using rationales for judging sarcasm. Appl Sci. 2024;14(11):4898. doi:10.3390/app14114898.
30. Prabhu R, Seethalakshmi V. A comprehensive framework for multi-modal hate speech detection in social media using deep learning. Sci Rep. 2025;15(1):13020. doi:10.1038/s41598-025-94069-z.
31. Sharma D, Nath T, Gupta V, Singh VK. Hate speech detection research in south Asian languages: a survey of tasks, datasets and methods. ACM Trans Asian Low-Resour Lang Inf Process. 2025;24(3):1–44. doi:10.1145/3711710.
32. Aïmeur E, Amri S, Brassel G. Fake news, disinformation and hate speech: defending the struggle for democracy in the digital age. Technologies. 2023;11(1):26. doi:10.3390/technologies11010026.
33. Nurfiqri MR, Fitriyani. The performance analysis of graph neural network (GNN) and convolutional neural network (CNN) algorithms for cyberbullying detection in twitter comments. IJCS. 2024;13(3):3656–57. doi:10.33022/ijcs.v13i3.3940.
34. Angger Saputra R, Sibaroni Y. Multilabel hate speech classification in Indonesian political discourse on X using combined deep learning models with considering sentence length. J Ilmu Komputer Dan Informasi. 2025;18(1):113–25. doi:10.21609/jiki.v18i1.1440.
35. Ibrohim MO, Budi I. A dataset and preliminaries study for abusive language detection in Indonesian social media. Procedia Comput Sci. 2018;135:222–9. doi:10.1016/j.procs.2018.08.169.
36. Alfina I, Mulia R, Fanany MI, Ekanata Y. Hate speech detection in the Indonesian language: a dataset and preliminary study. In: 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS); 2017 Oct 28–29; Bali, Indonesia. p. 233–8. doi:10.1109/ICACSIS.2017.8355039.
37. Mody D, Huang Y, Alves de Oliveira TE. A curated dataset for hate speech detection on social media text. Data Brief. 2022;46:108832. doi:10.1016/j.dib.2022.108832.
38. Almahdi AJ, Mohades A, Akbari M, Heidary S. Enhancing cross-lingual hate speech detection through contrastive and adversarial learning. Eng Appl Artif Intell. 2025;147(9):110296. doi:10.1016/j.engappai.2025.110296.
39. Alkomah F, Ma X. A literature review of textual hate speech detection methods and datasets. Information. 2022;13(6):273. doi:10.3390/info13060273.

40. Alhazmi A, Mahmud R, Idris N, Mohamed Abo ME, Eke CI. Code-mixing unveiled: enhancing the hate speech detection in Arabic dialect tweets using machine learning models. PLoS One. 2024;19(7):e0305657. doi:10.1371/journal.pone.0305657.
41. Capuano N, Fenza G, Loia V, Stanzione C. Explainable artificial intelligence in CyberSecurity: a survey. IEEE Access. 2022;10(2):93575–600. doi:10.1109/ACCESS.2022.3204171.
42. Madhu H, Satapara S, Modha S, Mandl T, Majumder P. Detecting offensive speech in conversational code-mixed dialogue on social media: a contextual dataset and benchmark experiments. Expert Syst Appl. 2023;215(1):119342. doi:10.1016/j.eswa.2022.119342.
43. Bhattacharya M, Roy S, Chattopadhyay S, Das AK, Shetty S. A comprehensive survey on online social networks security and privacy issues: threats, machine learning-based solutions, and open challenges. Secur Priv. 2023;6(1):e275. doi:10.1002/spy2.275.
44. Muzakir A, Adi K, Kusumaningrum R. Classification of hate speech language detection on social media: preliminary study for improvement. In: Emerging trends in intelligent systems & network security. Cham, Switzerland: Springer International Publishing; 2022. p. 146–56. doi:10.1007/978-3-031-15191-0_14.
45. Hakim AN, Sibaroni Y, Prasetyowati SS. Detection of hate-speech text on Indonesian twitter social media using IndoBERTweet-BiLSTM-CNN. In: 2024 12th International Conference on Information and Communication Technology (ICoICT); 2024 Aug 7–8; Bandung, Indonesia. p. 374–81. doi:10.1109/ICoICT61617.2024.10698615.
46. Santosa H, Rachman F, Austen SA, Christianto, Girsang AS. IndoBERT for classifying hate speech in Twitter. In: AIP Conference Proceedings. Chicago, IL, USA: AIP Publishing LLC; 2024. p. 050015. doi:10.1063/5.0199750.
47. Susanto L, Wijanarko MI, Pratama PA, Hong T, Idris I, Aji AF, et al. IndoToxic2024: a demographically-enriched dataset of hate speech and toxicity types for indonesian language. arXiv:2406.19349. 2024.