Indonesian Journal of Computational Language Studies
Vol. 1, No. 3, September 2026, pp. 178~191
ISSN: 3163-7884, DOI: https://doi.org/10.67490/ijcl.v1i3.1028

Detecting harm in context: toxicity, politeness, and cultural pragmatics in Indonesian social media NLP

M. Rifa Jamaludin Nasir
Universitas Islam Negeri Salatiga, Indonesia

Article Info

Article history:

Received Jul 22, 2026
Revised Aug 19, 2026
Accepted Sep 25, 2026

Keywords:

annotation uncertainty;
cultural pragmatics;
Indonesian NLP;
politeness;
toxicity detection.

ABSTRACT

Background: Indonesian social media toxicity detection remains difficult because harmful meaning is often produced through profanity, indirectness, ridicule, code-mixing, identity reference, and culturally situated politeness rather than through abusive vocabulary alone. **Objective:** This study examines how toxicity, politeness, and cultural pragmatics interact in Indonesian social media NLP annotation and asks what contextual evidence is required to distinguish abuse, hate speech, satire, criticism, and banter. **Method:** Using 60 anonymised Indonesian Twitter/X items from a public multi-label hate speech and abusive-language corpus, this study applies a mixed computational-pragmatic design combining source labels with manual interpretation of target, severity, facework, contextual validation, and risk. **Results:** Findings show that toxicity classes are evenly distributed, but target, domain, and severity differentiate surface abuse from identity-targeted harm. Facework cues such as direct address, laughter, ridicule, and abusive labels appear unevenly across toxicity classes, demonstrating that politeness and impoliteness cannot be inferred from lexical toxicity. Cultural-pragmatic coding shows that religion, ethnicity, gender, platform traces, political indexing, and local-language ambiguity require contextual validation before model outputs can be treated as reliable. **Implication:** Indonesian NLP moderation should combine classification with pragmatic annotation. **Novelty:** This study reframes toxicity detection as culturally grounded, target-aware, and pragmatically validated harm interpretation.

This is an open access article under the CC BY-SA license.

Corresponding Author:

M. Rifa Jamaludin Nasir
Universitas Islam Negeri Salatiga, Indonesia
Jl. Lingkar Salatiga Km. 2, Pulutan, Sidorejo, Kota Salatiga, Jawa Tengah, 50714, Indonesia
Email: rifa.jamaluddin@uinsalatiga.ac.id

1. INTRODUCTION

Indonesia offers a critical setting for examining toxicity detection because online harm now circulates within a very large, linguistically diverse, and culturally stratified digital public sphere. DataReportal [1] reports that Indonesia had 230 million internet users at the end of 2025, with internet penetration reaching 80.5%, and 180 million active social media user identities, equivalent to 62.9% of the national population. Within this scale, automated moderation cannot treat harmful language as a simple matter of profanity detection. Indonesian social media interaction often combines standard Indonesian, colloquial Indonesian, English, and regional languages, while insults, ridicule, moral accusation, religious insinuation, and sarcastic politeness may shift meaning according to target, relationship, platform, and topic. Recent work also reports a ten-fold increase in Indonesia's online hate-speech ratio over two years and introduces a 43,692-entry toxicity dataset annotated by 19 individuals. These facts make contextual harm detection a methodological necessity.

Prior research has built important foundations for Indonesian NLP, but it has not yet fully connected toxicity, politeness, and cultural pragmatics within one analytic model. Earlier Indonesian hate-speech detection already moved beyond binary classification by detecting target, category, and severity in Twitter data. More recent Indonesian computational studies have expanded technical coverage through named entity recognition, transformer-based political sentiment analysis, Indonesian–English code-mixing detection, lexical pattern analysis, readability assessment, and low-resource dialect segmentation [2], [3], [4], [5], [6], [7]. At the same time, culturally grounded LLM evaluation has begun to show that Indonesian language understanding cannot be reduced to translated benchmarks, especially when local languages, safety-sensitive prompts, factual-cultural errors, and sociolinguistic bias are involved [8], [9], [10], [11], [12]. Yet existing work still leaves underexplored how apparently polite or indirect utterances may function as harm. Sap et al. [24] show that hate-speech classifiers can disproportionately flag dialect-marked language, which illustrates how surface cues can misread social context.

This study addresses that gap by asking how harmful meaning is produced, softened, disguised, or intensified in Indonesian social media discourse when toxicity is read together with politeness and cultural pragmatics. More specifically, this study examines three linked questions: how explicit toxicity is distributed across surface forms, targets, and severity; how politeness and impoliteness strategies complicate annotation decisions; and how cultural cues such as hierarchy, religious reference, ethnicity, regional-language indexing, code-mixing, sarcasm, and indirectness shape contextual interpretation. This study therefore does not ask whether a model can merely classify posts as toxic or non-toxic. It asks what kinds of contextual information are necessary for Indonesian NLP systems and annotation protocols to distinguish abuse from criticism, protest, satire, in-group joking, or culturally coded hostility. Methodologically, this requires combining computational annotation with pragmatic interpretation, so that model-readable labels are evaluated against discourse-level evidence rather than isolated lexical triggers.

This study advances a provisional argument: Indonesian toxicity detection becomes more reliable and socially accountable when harm is treated as a contextual pragmatic act, not only as a textual attribute. A post may contain no overt profanity yet still attack dignity through mock respect, insinuation, exclusionary address, or stereotype activation; conversely, a direct or emotionally charged post may be wrongly penalized if its pragmatic function is protest, solidarity, or counter-speech. This argument aligns with emerging multilingual safety research showing that Indonesian-centric models remain vulnerable in colloquial and local-language settings, while culturally grounded safety evaluation is needed for responsible deployment. It also extends Indonesian NLP research on code-mixing, low-resource languages, safety failure, and sociolinguistic bias by placing annotation uncertainty at the center of analysis [4], [13], [10], [12]. This study consequently tests whether contextual coding can expose misclassification risks hidden by surface-level performance metrics. Mostafazadeh Davani et al. [25] argue that annotator disagreement in subjective tasks should be modeled rather than collapsed into a majority label, which supports treating annotation uncertainty as analytically meaningful.

2. LITERATURE REVIEW

2.1. Toxicity in social media NLP

Toxicity in social media NLP is usually defined as language likely to harm, exclude, threaten, or degrade another person or group, yet this definition remains unstable because harm may be lexical, affective, relational, or contextual. In computational work, toxicity often becomes a label attached to profanity, hate speech, abuse, or offensive sentiment, as seen in Indonesian research on political sentiment and transformer-based classification [3]. However, pragmatic theory suggests that harmful meaning is not exhausted by surface wording, because utterances perform social actions within situations [14], [15]. This distinction is crucial for Indonesian NLP: toxicity must be read as discourse practice, not merely as text classification.

The main indicators of toxicity include explicit abuse, targeted hostility, stereotyping, threats, dehumanization, humiliation, and severity. Indonesian NLP studies have improved technical identification of linguistic units, including named entities, code-mixed forms, lexical patterns, readability, and dialectal segmentation [2], [4], [5], [6], [7]. These works show that reliable classification depends on how language is segmented, normalized, and represented before modeling. Yet toxicity indicators remain methodologically fragile when abusive meaning is displaced into metaphor, sarcasm, indirect address, religious insinuation, or local-language indexing. This study therefore treats toxicity as a layered phenomenon involving target, surface form, severity, and contextual ambiguity.

2.2. Politeness

Politeness is commonly understood as linguistic management of face, respect, distance, and social relation, but scholars differ on whether politeness is a universal strategic system or an interactionally negotiated judgement. Brown and Levinson [16] conceptualize politeness as face-saving strategy, while Locher and Watts [17] argue that politeness emerges from situated evaluations of appropriate conduct.

Spencer-Oatey [18] further shifts attention to rapport, rights, obligations, and sociality. In Indonesian social media, this distinction matters because formally respectful address may still wound, mock, discipline, or exclude. Politeness is therefore not simply opposed to toxicity; it can become a vehicle for subtle harm.

Indicators of politeness and impoliteness include honorifics, indirectness, mitigation, apology, address terms, mock respect, sarcasm, ridicule, dismissiveness, and face attack. Culpeper's work on impoliteness [19], [20] is useful because it explains how aggression may operate through conventionalized strategies rather than overt insults alone. Grice's [21] theory of implicature also helps explain how speakers can communicate hostile meanings indirectly, allowing plausible deniability. Indonesian computational studies on code-mixing, lexical patterning, and educational readability provide tools for detecting linguistic variation, but they do not by themselves explain relational force [4], [5], [6]. This study links these indicators to annotation uncertainty.

2.3. Cultural pragmatics

Cultural pragmatics refers to how meaning is shaped by shared norms, local histories, status relations, religious references, ethnic categories, regional identities, and expectations of appropriateness. Unlike formal semantics, cultural pragmatics asks why a phrase becomes harmful in one social setting but ordinary, humorous, or solidaristic in another. Recent Indonesian-oriented LLM studies make this issue unavoidable by questioning translated benchmarks, multilingual reasoning consistency, culturally sensitive safety, factual-cultural hallucination, low-resource adaptation, and sociolinguistic bias [8], [9], [10], [11], [13], [12]. Their shared implication is that Indonesian NLP evaluation must move beyond standard-language assumptions.

Indicators of cultural pragmatics include code-mixing, regional-language markers, religious or ethnic indexing, hierarchy-sensitive address, dog whistles, stereotype activation, moral accusation, collective identity cues, and thread-level context. Haugh's [22] interactional approach to face supports this view because face is co-constructed across participants rather than stored in isolated words. Indonesian NLP work on dialect segmentation and code-mixing shows that models must handle multilingual variation, while bias and safety studies warn that underrepresented speakers may be misread or underprotected [7], [4], [10], [12]. This study positions toxicity detection as contextual pragmatic interpretation supported by computation, not replaced by it.

3. METHOD

This study uses a tweet-level unit of analysis to examine how harmful meaning is encoded in Indonesian social media text through toxicity, politeness, and cultural pragmatics. Primary data consist of 60 anonymised Indonesian Twitter/X items drawn from a public multi-label hate speech and abusive-language corpus. Six additional records are used as secondary and contextual sources to support dataset provenance, multilingual relevance, and methodological alignment. Usernames and URLs remain anonymised, while abusive expressions are sanitised for ethical reporting. Because original tweet dates are not provided in the public release, temporal claims are limited to dataset release and source-document dates, not posting chronology.

This study adopts a mixed computational-pragmatic design. Computationally, source labels are used to organise toxicity, abusive language, target scope, target domain, and severity without redefining original categories. Pragmatically, each text is prepared for manual interpretation of facework, impoliteness, indirect harm, code-mixing, and contextual ambiguity. This design follows Indonesian NLP research that treats language processing as label-sensitive and domain-specific, including work on sentiment analysis, code-mixing detection, dialect segmentation, named entity recognition, and culturally grounded language-model evaluation [3], [4], [7], [2], [8], [10]. This design supports descriptive and interpretive claims, not causal inference.

Information sources are divided into primary, secondary, and contextual materials. Primary sources provide anonymised social media text and original multi-label annotations. Secondary sources document dataset provenance, label logic, and comparable Indonesian toxicity resources. Contextual sources support interpretation of Indonesian multilingual NLP, local-language safety, cultural bias, and sociolinguistic representation, particularly in relation to LLM evaluation and low-resource Indonesian languages [9], [11], [13], [12]. This separation prevents source conflation: primary text supplies analyzable discourse, secondary sources support methodological validation, and contextual sources frame interpretation without being treated as empirical tweet-level evidence.

Table 1. Composition of dataset or sample

Code	Data Type	Source	Period	Quantity	Characteristics
PRM	Primary corpus	Public Indonesian Twitter/X hate speech and abusive-language dataset	Public release: 2019; original tweet dates unavailable	60	Anonymised tweet-level items with source labels for hate speech, abusive language, target, category, and severity
SEC	Secondary source	Peer-reviewed paper and comparator dataset cards	2019–2025	2	Used to verify label architecture and compare dataset scope
CTX	Context source	Dataset cards and literature on Indonesian toxicity, local languages, and safety evaluation	2021–2025	4	Used to contextualise code-mixing, cultural safety, and multilingual evaluation
DEDUP	Deduplication frame	Original primary CSV	Public release: 2019	13,169 original rows; 155 duplicates removed; 13,014 unique rows	Normalised-text hash deduplication before purposive selection
VIS	Visual/video fields	Corpus workbook schema	Not applicable	0 visual/video items	Image description, actor, symbol, colour, layout, logo position, frame, timestamp, transcript, and screenshot fields retained as not applicable

Data collection proceeded through public-source identification, keyword mapping, source registration, metadata capture, text storage, sanitisation, and deduplication. Search terms included “Indonesian hate speech,” “abusive language,” “toxicity,” “politeness,” “Indonesian Twitter,” “code-mixing,” “cultural pragmatics,” and “LLM safety Indonesian languages.” Each item was assigned a unique document code and recorded with title, date status, source, link, context, data role, language scope, inclusion status, exclusion status, and ethical note. Duplicate texts were removed through normalised-text hashing before analytic selection. Visual and video fields were preserved in this workbook schema, but all current items are text-only.

Data analysis proceeds through three linked instruments. First, a toxicity matrix maps source labels into surface harm, target scope, target category, severity, and review priority. Second, a politeness–impoliteness grid examines address, ridicule, mock respect, mitigation, sarcasm, and face-threatening cues. Third, a cultural-pragmatic protocol assesses religious, ethnic, political, regional-language, code-mixing, and thread-dependent ambiguity. These instruments convert dataset labels into analytically interpretable categories while retaining boundaries between source annotation and researcher interpretation. Manual validation remains necessary before final claims about pragmatic harm, because source toxicity labels alone cannot establish politeness status, cultural meaning, or misclassification risk.

Table 2. Operational matrix of analysis

Analytical Instrument	Main Question	Input Variables	Coding Operation	Claim Boundary
Toxicity surface, target, and severity matrix	What forms of source-coded harm appear in Indonesian social media text?	Hate speech, abusive language, individual/group target, category, weak/moderate/strong severity	Map source labels into descriptive toxicity classes and review priority	Descriptive only; no causal claim
Politeness–impoliteness and facework grid	How may harm be softened, intensified, or disguised through pragmatic form?	Sanitised text, address terms, laughter markers, mitigation, mock respect, sarcasm, directness	Manual pragmatic coding of face-threatening and facework signals	Requires coder validation
Cultural pragmatics and contextual ambiguity protocol	Which Indonesian sociocultural cues make harm detection context-dependent?	Target domain, religious/ethnic/political cues, code-mixing, contextual need, ambiguity	Identify culturally situated cues and misclassification risk	Interpretive claims must remain evidence-bound

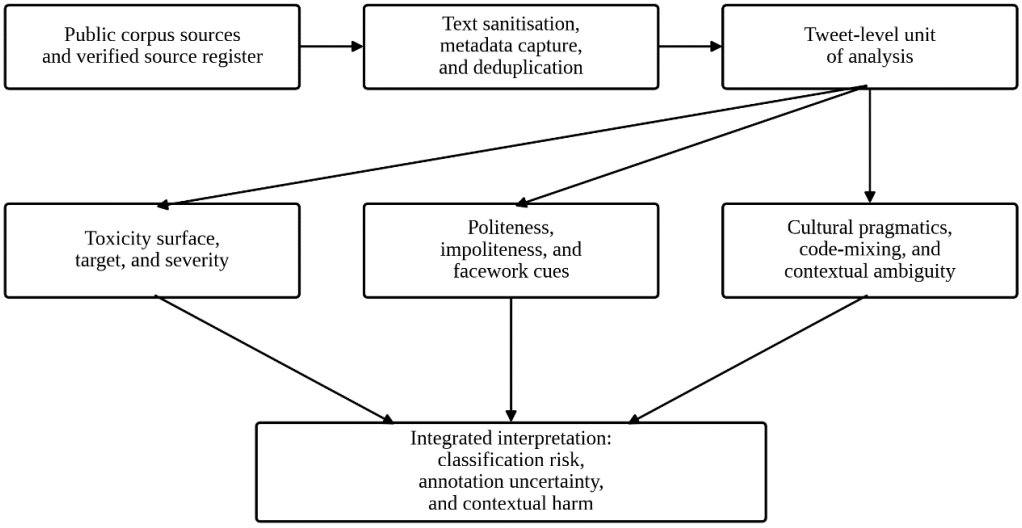

Figure 1. Operational matrix of analysis

4. RESULTS

4.1. Toxicity surfaces, targeting, and severity in Indonesian social media text

The matrix shows that harmful-language detection in this corpus cannot be reduced to a single binary distinction between toxic and non-toxic expression. Four source-coded surface classes are equally represented, but their internal severity and target structures differ sharply. Hate speech combined with abusive language is concentrated in weak and moderate levels, while hate speech without abusive-language marking includes two strong-severity cases. Abusive-only items and control items carry no source-coded hate target, domain, or severity. This distribution matters methodologically because a classifier trained only to detect abusive vocabulary would not capture how target, domain, and severity reorganise harm. Indonesian NLP work on sentiment, code-mixing, and low-resource processing similarly suggests that reliable classification depends on how linguistic units are operationalised before modelling [3], [4], [7].

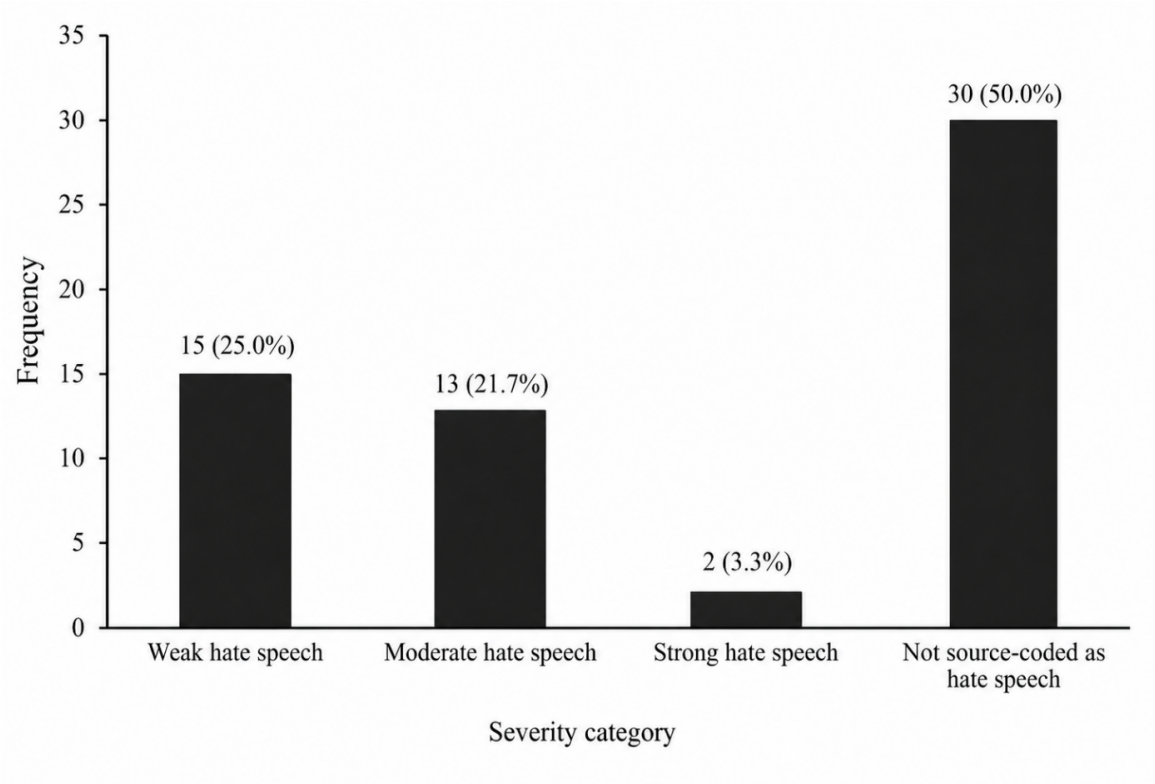

Figure 2. Source-coded severity distribution in Indonesian social media text

Dominant patterns emerge around the separation between surface abuse and identity-targeted harm. Half of the corpus has no source-coded hate target, while 16 items target individuals and 14 target groups. The largest hate domain is "other invective/slander" with 18 items, followed by smaller but analytically important categories related to religion, race/ethnicity, gender/sexuality, and one mixed religion–physical/disability label. The excerpts illustrate this difference: "*endingnya … [ABUSIVE_TERM] … friendzone*" functions as abusive-only evaluation, whereas "*[ETHNIC_GROUP] … usir*" is source-coded as strong hate speech because it activates group-based exclusion. Similarly, "*kaum [GROUP_LABEL] … hahahah*" combines group reference with ridicule. These patterns indicate that severity depends not only on offensive wording, but also on targetability and social category.

The pattern supports a pragmatic view of toxicity: harmfulness becomes analytically visible when surface form, target scope, social domain, and severity are read together. A direct insult toward an individual, a joking abusive expression, and a group-directed exclusionary utterance may all contain hostile language, but they do not carry the same annotation consequence. This distinction is consistent with Indonesian computational studies that emphasise culturally grounded language understanding, multilingual reasoning, safety sensitivity, factual-cultural error, and sociolinguistic bias [8], [9], [10], [11], [12]. This study therefore treats toxicity detection as a layered interpretive task. The provisional implication is that Indonesian social media NLP requires target-aware and context-sensitive annotation before model performance can be meaningfully evaluated.

Table 3. Toxicity surface, target, domain, and severity matrix

Analytical Dimension	Main Category	Operational Indicator	Frequency	Percentage
Toxicity surface	Hate speech + abusive language	Source-coded as both hate speech and abusive language	15	25.0%
Toxicity surface	Hate speech only	Source-coded as hate speech without abusive-language label	15	25.0%
Toxicity surface	Abusive language only	Source-coded as abusive but not hate speech	15	25.0%
Toxicity surface	Non-harm / control label	Not source-coded as hate speech or abusive language	15	25.0%
Target scope	Individual target	Harm directed toward an identifiable person or personal addressee	16	26.7%
Target scope	Group target	Harm directed toward a collective identity or social group	14	23.3%
Target scope	No source-coded hate target	No hate-speech target in source annotation	30	50.0%
Target domain	Other invective / slander	General insult, political invective, or slander category	18	30.0%
Target domain	Religion / creed	Religion-related target category	4	6.7%
Target domain	Race / ethnicity	Ethnic or racialised target category	4	6.7%
Target domain	Gender / sexuality	Gendered or sexuality-related target category	3	5.0%
Target domain	Religion / creed + physical / disability	Combined source-coded domain	1	1.7%
Target domain	No source-coded domain	No hate-speech domain assigned	30	50.0%
Severity	Weak hate speech	Source-coded weak hate-speech severity	15	25.0%
Severity	Moderate hate speech	Source-coded moderate hate-speech severity	13	21.7%
Severity	Strong hate speech	Source-coded strong hate-speech severity	2	3.3%
Severity	Not source-coded as hate speech	No hate-speech severity label assigned	30	50.0%

4.2. Politeness, impoliteness, and facework cues in harm annotation

Indonesian harmful-language annotation becomes more complex once facework is placed beside toxicity labels. The corpus keeps the same 60 primary tweet-level items used in the previous analysis, but shifts attention from source-coded toxicity to interactional cues. The largest category is the absence of salient facework cues in automatic screening, covering 26 items, followed by abusive items without explicit politeness coding, covering 20 items. Direct address appears in 11 items, while ridicule or laughter markers appear in four items, including one combined with direct address. This distribution shows that politeness and impoliteness cannot be inferred directly from hate-speech or abusive-language labels. Work on Indonesian code-mixing and culturally grounded evaluation similarly suggests that textual classification requires pragmatic interpretation, not only lexical recognition [4], [8], [10].

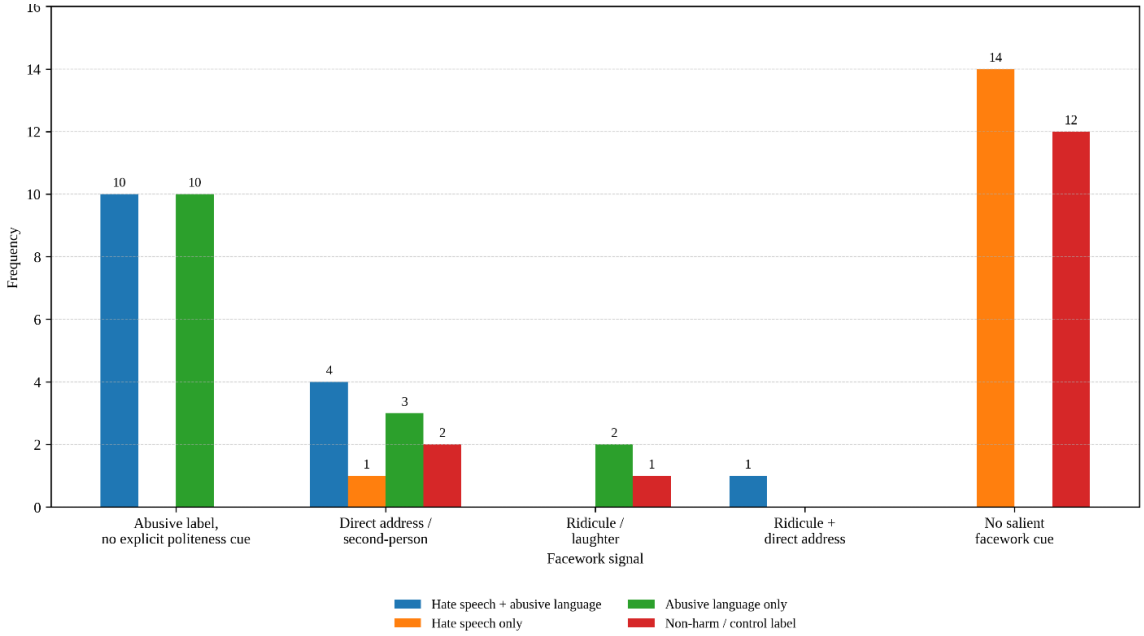

Figure 3. Facework signal by toxicity class

The dominant pattern is that explicit facework cues are unevenly attached to toxicity classes. Abusive-source items are concentrated in two places: 10 appear as hate speech plus abusive language, and 10 appear as abusive-only language. Direct address crosses all four surface classes, which means addressee orientation is not automatically toxic; it can appear in control items as well as harmful ones. Ridicule markers are similarly ambiguous: "*wkwk*" appears in abusive-only and control items, while "*hahahah*" in "*Kaum [ABUSIVE_TERM] ... hahahah*" intensifies a group-directed utterance. By contrast, "*PKI hanya muncul jika jelang pemilihan ... digoreng sampai gosong*" has no overt direct-address marker but remains source-coded through political and ideological reference. The corpus therefore separates interactional form from toxicity status rather than treating them as equivalent.

Facework analysis reframes harm as an interactional achievement rather than a fixed property of offensive vocabulary. Brown and Levinson's [16] account of face, Culpeper's [19], [20] model of impoliteness, and Locher and Watts' [17] interactional view of relational judgement help explain why direct address, laughter, mockery, and apparent casualness may alter annotation risk. In Indonesian NLP, this issue is intensified by informal particles, platform-specific laughter, code-mixing, ideological labels, and culturally recognisable address forms. This finding also connects with recent Indonesian studies on multilingual reasoning, local-language safety, factual-cultural error, and sociolinguistic bias [9], [11], [12]. This study therefore positions politeness not as a separate layer added after toxicity detection, but as a pragmatic condition that shapes whether harm is visible, softened, intensified, or misclassified.

Table 4. Facework signal, toxicity class, and pragmatic annotation matrix

Analytical Dimension	Facework Category	Operational Indicator	Frequency	Percentage	Distribution across Toxicity Class
Facework visibility	Abusive source label without explicit politeness coding	Item is source-coded as abusive, but no explicit politeness marker is identified in screening	20	33.3%	Hate speech + abusive language = 10; abusive language only = 10
Address orientation	Direct address / second-person orientation	Use of "*loe*," "lu," "*kau*," "*kamu*," or interactional orientation toward an addressee	11	18.3%	Hate speech + abusive language = 5; hate speech only = 1; abusive language only = 3; control = 2
Ridicule marker	Ridicule / laughter marker	Laughter, mockery, playful ridicule, or derisive affect markers such as "*hahahah*" or "*wkwk*"	4	6.7%	Hate speech + abusive language = 1; abusive language only = 2; control = 1
Combined facework	Ridicule plus direct address	Ridicule marker occurs together with second-person or addressee-oriented expression	1	1.7%	Hate speech + abusive language = 1
Low explicit facework visibility	No salient facework cue in automatic screening	No direct address, ridicule marker, or explicit politeness/impoliteness cue is identified	26	43.3%	Hate speech only = 14; control = 12

4.3. Cultural pragmatics, contextual validation, and misclassification risk

Cultural-pragmatic evidence shows that contextual validation is not an optional supplement to toxicity detection but a condition for defensible annotation. The same 60 tweet-level items remain the unit of analysis, while this layer asks whether harm can be interpreted from text alone or requires knowledge of platform form, target identity, political discourse, religious reference, ethnicity, gender, or local-language variation. Thirty-five items have no explicit cultural-pragmatic signal in source labels, yet this does not mean they are pragmatically simple. Twenty of those still require medium thread or topic validation. Every item is also flagged for manual code-mixing or local-language review because source labels do not encode multilingual status. This pattern supports recent Indonesian NLP work on low-resource language adaptation, cultural safety, and sociolinguistic bias.

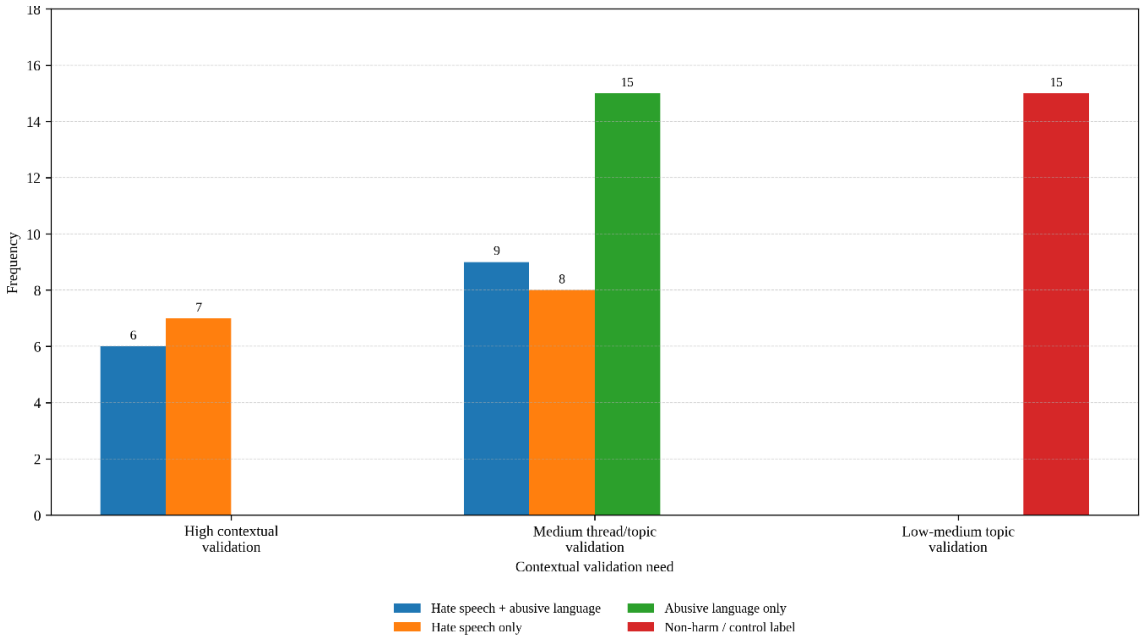

Figure 4. Contextual validation need by toxicity class

The dominant pattern is a split between explicit identity-sensitive harm and context-dependent non-identity discourse. High contextual validation appears in 13 items: six hate speech plus abusive-language items and seven hate-speech-only items. These include religion, ethnicity, gender/sexuality, and mixed religion–physical/disability categories. By contrast, medium validation dominates with 32 items, including all 15 abusive-only cases and 17 hate-speech-coded cases whose interpretation depends on thread, topic, or political context. Low-medium validation is restricted to the 15 control items. The excerpts clarify this distinction. "*[ETHNIC_GROUP] perusak bangsa!! Usir!!*" requires high validation because exclusion is attached to ethnic identity, while "USER USER *masa baru kena hestek aja udah kepanasan*" requires topic validation because platform interaction and political hashtag context shape its force.

Cultural pragmatics explains why Indonesian toxicity detection cannot rely only on source labels or surface offensiveness. A model may correctly detect abusive terms but still miss whether harm is produced through religious indexing, ethnic exclusion, gendered norm enforcement, factional ridicule, or anonymised platform interaction. Conversely, a politically sharp or emotionally intense statement may be over-classified as harmful if its thread-level function is criticism, satire, or counter-speech. This position is consistent with Bouazzaoui's [8] culturally grounded evaluation, Wulandari's [9] multilingual reasoning concern, Nadiyah's [10] safety-oriented analysis of local languages, Hadziqi's [13] low-resource adaptation framework, and Sumadi's [12] work on sociolinguistic bias. This study therefore argues that Indonesian social media NLP requires cultural-pragmatic validation as a bridge between annotation labels, model prediction, and socially accountable interpretation.

Table 5. Cultural-pragmatic signal, contextual need, and misclassification risk matrix

Analytical Dimension	Cultural-Pragmatic Category	Operational Indicator	Frequency	Percentage	Contextual Validation Pattern	Misclassification Risk Pattern
Source-label limitation	No cultural-pragmatic signal coded from source labels	No religion, ethnicity, gender, platform token, or political indexing explicitly captured in cultural-pragmatic screening	35	58.3%	Medium = 20; low-medium = 14; high = 1	High = 15; medium = 6; low-medium = 14
Platform mediation	Platform-specific anonymised token present	USER, URL, repost markers, or anonymised platform traces affect interpretation	8	13.3%	Medium = 7; low-medium = 1	High = 7; low-medium = 1
Political indexing	Political faction slang or ideological indexing with platform token	Political slogans, ideological labels, campaign hashtags, or factional references occur with anonymised platform traces	4	6.7%	Medium = 4	High = 2; medium = 2
Race / ethnicity	Race or ethnicity category in source label	Source-coded race/ethnicity target without additional platform token	3	5.0%	High = 3	High = 3
Religion and platform	Religion/creed category with platform token	Religion-related target category appears with anonymised user or repost trace	2	3.3%	High = 2	High = 1; medium = 1
Religion, politics, and platform	Religion/creed plus political indexing and platform token	Religious category intersects with political or ideological framing	2	3.3%	High = 2	High = 1; medium = 1
Gender / sexuality	Gender or sexuality category in source label	Gendered or sexuality-related target category	2	3.3%	High = 2	Medium = 2
Gender and platform	Gender/sexuality category with platform token	Gendered target category appears with anonymised platform trace	1	1.7%	High = 1	Medium = 1
Religion and physical/disability	Religion/creed plus physical/disability category	Source label combines religious reference and bodily/disability-related insult	1	1.7%	High = 1	Medium = 1
Political indexing only	Political faction slang or ideological indexing without platform token	Political or ideological reference appears without explicit platform token	1	1.7%	Medium = 1	Medium = 1
Race / ethnicity and platform	Race/ethnicity category with platform token	Ethnic/racial category appears with anonymised platform trace	1	1.7%	High = 1	High = 1
Manual review requirement	Code-mixing/local-language review	Source labels do not identify code-mixing or local-language status	60	100.0%	All items require manual review for code-mixing/local-language ambiguity	Not used as independent risk score

5. DISCUSSION

The first analytic layer shows that toxicity detection in Indonesian social media becomes dysfunctional when harm is treated as a single textual property. Equal distribution across hate speech plus abuse, hate speech only, abusive-only language, and control items does not imply equal pragmatic risk; rather, it shows that surface categories organise harm differently. This matters because abusive vocabulary can appear without identity-targeted hate, while hate speech may appear without overt abusive markers. The implication is methodological: Indonesian NLP systems that privilege profanity or sentiment intensity risk over-detecting rude expression and under-detecting targeted exclusion. This finding extends Ibrohim and Budi's [23] multi-label approach by affirming that target, category, and severity remain indispensable to annotation. It also qualifies transformer-based optimism in Indonesian toxicity detection: higher classification capacity does not automatically solve harm interpretation. This study therefore argues that toxicity labels should function as structured evidence, not final explanations of social harm.

The structure behind this pattern lies in the unequal relationship between surface offensiveness and social targetability. Abusive language is often lexically visible, but hate speech depends on whether an utterance attaches negativity to a person, group, or protected identity category. This explains why strong severity appears only in a small subset, while non-hate or abusive-only cases occupy much larger space. The finding supports Santoso's [2] emphasis on entity-sensitive language processing and Nugroho's [4] attention to code-mixing because target recognition and linguistic hybridity shape what a model can identify. At the same time, it complicates Novitasari's [3] transformer-based political sentiment orientation: affective polarity alone cannot distinguish insult, critique, ideological antagonism, and exclusionary harm. This study does not claim causal relations between language form and harm level; rather, it shows that severity is structurally associated with target construction, category activation, and annotation scope. Harm detection must therefore begin with relational structure.

Facework evidence shows that politeness and impoliteness are not peripheral to toxicity detection but central to how harm becomes interactionally legible. Direct address, laughter markers, ridicule, and abusive labels do not merely decorate harmful language; they help shape whether an utterance functions as attack, mockery, dismissal, provocation, or ordinary platform banter. The dysfunction appears when moderation systems treat such cues as fixed signals. A second-person form may intensify an attack, but it may also occur in non-harmful interaction. Laughter may signal solidarity, ridicule, or contempt. This ambiguity supports Brown and Levinson's [16] view that face is interactionally managed, while also aligning more closely with Locher and Watts [17], who argue that politeness is judged within situated norms. For Indonesian NLP, this means politeness cannot be operationalised as a simple positive feature opposed to toxicity. This study positions facework as a risk-sensitive interpretive layer within annotation design.

This pattern emerges because social media language compresses interpersonal relation, stance, and audience performance into short textual forms. Indonesian digital interaction often uses casual address, particles, joking formats, laughter tokens, and mock familiarity, so the same formal feature can carry different pragmatic force across contexts. Culpeper's [19], [20] model of impoliteness helps explain why ridicule and mock respect may become face attack even when explicit threat is absent. Gricean implicature [21] also clarifies how speakers can imply hostility without saying it directly, leaving annotation teams to infer pragmatic force from incomplete evidence. This partly challenges models that rely on lexical patterning or readability-oriented segmentation, as in Wahid [5] and Yuliyana [6], because recognisable words do not necessarily disclose relational meaning. Yet it also complements their work by showing why preprocessing must be followed by pragmatic validation. The underlying structure is therefore interactional: harm is produced through relation, uptake, and socially recognisable cues.

Cultural-pragmatic evidence carries the strongest implication for Indonesian NLP safety: contextual validation is necessary whenever harm intersects with religion, ethnicity, gender, political identity, local-language forms, or platform mediation. Items without explicit cultural-pragmatic labels are not automatically simple, because absence of source-coded cultural signal may reflect annotation limits rather than absence of cultural meaning. This finding resonates with Bouazzaoui's [8] critique of translated benchmarks and Nadiyah's [10] concern with culturally sensitive safety failure in Indonesian languages. It also supports Sumadi's [12] argument that sociolinguistic bias shapes who is represented and protected by Indonesian AI systems. The practical implication is that benchmark performance may conceal culturally consequential false positives and false negatives. A model may detect an abusive term while missing ethnic exclusion, or penalise sharp political critique as toxicity. This study therefore treats cultural pragmatics as a safeguard against both under-protection and over-moderation.

This contextual dependency arises from Indonesia's multilingual and socially stratified communicative ecology. Harm is not always carried by universal slurs; it may be embedded in local-language

fragments, religious indexing, ideological shorthand, ethnic stereotypes, gendered norm enforcement, or platform-specific interactional traces. Ibrahim's [7] work on low-resource dialect segmentation and Hadziqi's [13] work on parameter-efficient multilingual adaptation are relevant because culturally situated toxicity often appears where models have weaker linguistic coverage. Wulandari's [9] multilingual evaluation and Wibisono's [11] analysis of factual-cultural error further suggest that Indonesian language understanding fails when models treat Indonesian as a single stable linguistic space. This does not mean every culturally marked utterance is harmful. Instead, this study shows why annotation must separate cultural cue, target, pragmatic force, and severity. The deeper structure is sociolinguistic: Indonesian harm detection depends on whose language is recognised, whose context is available, and whose vulnerability is made machine-readable. Fawaid [26] frames language, media, and representation as a politics of who is positioned to speak and be heard, which is consistent with this sociolinguistic reading of harm.

6. CONCLUSION

This study demonstrates that harm detection in Indonesian social media NLP becomes analytically stronger when toxicity is examined alongside politeness and cultural pragmatics. Its most important lesson is that harmful meaning cannot be reduced to abusive vocabulary, sentiment polarity, or binary toxicity labels. Surface abuse, identity-targeted hostility, face-threatening interaction, ridicule, indirectness, code-mixing, and culturally marked references operate through different pragmatic mechanisms. This study contributes by renewing toxicity detection as a context-sensitive annotation problem rather than a purely classificatory task. Its methodological strength lies in integrating target, severity, facework, and cultural validation within one coherent corpus-based framework for Indonesian social media analysis.

This study remains limited by its reliance on anonymised public corpus items whose original posting dates, full interactional threads, user relations, and platform-specific uptake are unavailable. These constraints restrict claims about temporal dynamics, conversational intention, audience reception, and model performance across platforms. Future research should expand this design through larger multilingual corpora, thread-level annotation, cross-platform comparison, annotator-agreement testing, and direct evaluation of transformer or LLM-based moderation systems. Further studies should also include Indonesian regional languages more systematically, especially where local pragmatic cues shape whether utterances are interpreted as abuse, satire, solidarity, political criticism, or culturally coded hostility.

ACKNOWLEDGMENTS

The author thanks the reviewers and editorial team for their constructive feedback on this manuscript.

FUNDING INFORMATION

Author states no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

M. Rifa Jamaludin Nasir: conceptualization, methodology, formal analysis, writing – original draft, writing – review and editing.

CONFLICT OF INTEREST STATEMENT

Author states no conflict of interest.

AI USE DECLARATION

The author used artificial intelligence tools to assist with the generation of illustrative figures and with the formatting and layout of this manuscript. AI was not used to generate the substantive content, conceptualization, data, analysis, or conclusions of the research, all of which remain the sole responsibility of the author.

INFORMED CONSENT

Not applicable – this study did not involve direct human participants. Data consisted solely of publicly accessible, anonymized public digital text records.

ETHICAL APPROVAL

This research did not involve human or animal subjects; it analyzed publicly accessible digital text records and scholarly materials, and therefore no institutional review board approval was required. Where research does involve human subjects, the author affirms that such research would be conducted in full compliance with all relevant national regulations and institutional policies, in accordance with the tenets of

the Declaration of Helsinki, and would be approved by the author's institutional review board or equivalent committee.

DATA AVAILABILITY

The corpus of publicly accessible digital text records and scholarly materials analyzed in this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] DataReportal, *Digital 2026: Indonesia*, Nov. 5, 2025. [Online]. Available: https://datareportal.com/reports/digital-2026-indonesia.

[2] J. Santoso, "Indonesian named entity recognition using BiLSTM-CRF with domain-specific word embeddings," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 1, pp. 1-10, 2026.

[3] W. Novitasari, "Sentiment analysis on Indonesian political discourse using transformer-based models," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 1, pp. 11-21, 2026.

[4] A. Nugroho, "Code-mixing detection in Indonesian-English tweets using machine learning and linguistic features," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 1, pp. 22-31, 2026.

[5] A. Wahid, "Lexical patterns in Indonesian academic writing: A corpus-based analysis for writing assistance," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 1, pp. 32-42, 2026.

[6] T. Yuliyana, "Automatic readability assessment of Indonesian educational texts using hybrid NLP approaches," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 1, pp. 43-52, 2026.

[7] M. I. Ibrahim, "Morphological segmentation of low-resource Indonesian dialects using unsupervised neural models," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 1, pp. 53-62, 2026.

[8] A. E. Bouazzaoui, "Beyond translated benchmarks: A culturally grounded evaluation of large language models for Indonesian language understanding," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 2, 2026.

[9] F. D. Wulandari, "Do large language models reason consistently across Indonesian languages? A multilingual evaluation of Indonesian, Javanese, Sundanese, and Buginese," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 2, 2026.

[10] Nadiyah, "When safety fails in local languages: Evaluating culturally sensitive responses of large language models in multilingual Indonesia," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 2, 2026.

[11] R. T. Wibisono, "Hallucinating the archipelago: Detecting factual and cultural errors in retrieval-augmented generation for Indonesian knowledge," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 2, 2026.

[12] M. A. Sumadi, "Who gets represented by Indonesian AI? Measuring regional, gender, and sociolinguistic bias in large language models," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 2, 2026.

[13] A. Hadziqi, "Small models, many languages: Parameter-efficient adaptation of multilingual language models for low-resource Indonesian languages," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 2, 2026.

[14] J. L. Austin, *How to Do Things with Words*. Oxford, U.K.: Oxford Univ. Press, 1962.

[15] J. R. Searle, *Speech Acts: An Essay in the Philosophy of Language*. Cambridge, U.K.: Cambridge Univ. Press, 1969.

[16] P. Brown and S. C. Levinson, *Politeness: Some Universals in Language Usage*. Cambridge, U.K.: Cambridge Univ. Press, 1987.

[17] M. A. Locher and R. J. Watts, "Politeness theory and relational work," *Journal of Politeness Research*, vol. 1, no. 1, pp. 9-33, 2005.

[18] H. Spencer-Oatey, "Face, im/politeness and rapport," in *Culturally Speaking: Culture, Communication and Politeness Theory*, 2nd ed., H. Spencer-Oatey, Ed. London: Continuum, 2008, pp. 11-47.

[19] J. Culpeper, "Towards an anatomy of impoliteness," *Journal of Pragmatics*, vol. 25, no. 3, pp. 349-367, 1996.

[20] J. Culpeper, *Impoliteness: Using Language to Cause Offence*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[21] H. P. Grice, "Logic and conversation," in *Syntax and Semantics: Speech Acts*, vol. 3, P. Cole and J. L. Morgan, Eds. New York: Academic Press, 1975, pp. 41-58.

[22] M. Haugh, "Im/politeness, social practice and participation order," *Journal of Pragmatics*, vol. 58, pp. 52-72, 2013.

[23] M. O. Ibrohim and I. Budi, "Multi-label hate speech and abusive language detection in Indonesian Twitter," in *Proc. 3rd Workshop on Abusive Language Online*, 2019, pp. 46-57, doi: 10.18653/v1/W19-3506.

[24] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 1668-1678, doi: 10.18653/v1/P19-1163.

[25] A. Mostafazadeh Davani, M. Díaz, and V. Prabhakaran, "Dealing with disagreements: Looking beyond the majority vote in subjective annotations," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 92-110, 2022, doi: 10.1162/tacl_a_00449.

[26] A. Fawaid, *Who Controls the Narrative? Language, Media, and Politics of Representation*. Probolinggo, Indonesia: Pustaka Nurja, 2026.

BIOGRAPHY OF AUTHOR



M. Rifa Jamaludin Nasir is affiliated with Universitas Islam Negeri Salatiga, Indonesia, where his work engages computational pragmatics, toxicity and hate-speech detection, and the study of politeness, impoliteness, and cultural context in Indonesian social media NLP. His current research examines how facework cues, cultural-pragmatic signals, and contextual validation shape the reliability of automatic harm detection for Indonesian-language digital text. He can be contacted at email: rifa.jamaluddin@uinsalatiga.ac.id