Beyond Annotator Disagreement: Guideline-Induced Errors in Arabic Hate Speech Annotation

Wajdi Zaghouani
Northwestern University Qatar
wajdi.zaghouani@northwestern.edu

Abstract

Annotation errors in hate speech corpora are often attributed to annotator disagreement or bias. This paper argues that a substantial and underexamined class of errors originates upstream, from structural weaknesses in annotation guidelines themselves. When guidelines fail to encode the linguistic and cultural properties of the target discourse, they make certain errors structurally inevitable regardless of annotator quality. Focusing on Arabic social media discourse, a challenging setting due to its dialect continuum, culturally embedded insult conventions, sarcasm-heavy pragmatics, and complex religious rhetoric, we identify three mechanisms through which guideline design produces systematic annotation errors: cultural misclassification, when culturally specific hostile expressions fall outside annotation categories; dialectal ambiguity, when lexical meanings shift across regional varieties; and annotation projection, when frameworks developed for English moderation are applied to Arabic without adequate adaptation. Using six illustrative case studies with attested Arabic examples, we show how these mechanisms produce recurrent misannotations in existing datasets. We propose a taxonomy of five guideline-induced error types, an explicit mapping from mechanisms to error types, and a practical four-stage diagnostic framework for dataset builders.

1 Introduction

Annotation errors are a persistent challenge in NLP dataset construction, and their sources are not always well understood.

In the hate speech detection literature, annotation inconsistencies are most frequently attributed to individual annotators: their disagreements, biases, or insufficient training (Waseem and Hovy, 2016; Davidson et al., 2017). Inter-annotator agreement (IAA) metrics such as Cohen's κ and Krippendorff's α are routinely used to assess reliability (Artstein and Poesio, 2008). While such metrics are informative, recent work has emphasized that they conflate two distinct phenomena. Pavlick and Kwiatkowski (2019) demonstrate that annotator disagreement often reflects genuine linguistic uncertainty rather than annotation failure. Nie et al. (2020) further show through large-scale collection of multiple annotations per item that meaningful distributions of human opinion persist even on widely used NLI evaluation sets, and Jiang et al. (2023) provide ecologically valid explanations for why such variation arises. Weber-Genzel et al. (2024) introduce a rigorous methodology for separating annotation errors from human label variation, establishing that the two require different remediation strategies. Klie et al. (2023) survey annotation error detection methods across eighteen approaches and nine datasets, showing that many widely used corpora contain substantial numbers of objectively incorrect labels.

This paper contributes to that line of inquiry by examining a neglected *upstream* source of annotation error: the guidelines themselves. We argue that when annotation guidelines fail to account for the linguistic and cultural properties of the target discourse, they systematically produce errors regardless of annotator quality. These are not errors caused by annotator sloppiness or disagreement; they are errors *structurally forced* by the design of the annotation scheme, inscribed before a single label is assigned.

We develop this argument in the context of Arabic social media discourse, which is a particularly vulnerable setting for guideline-induced errors. Four properties compound the risk: (1) a *dialect continuum* in which the same word carries sharply different meanings across Gulf, Levantine, Egyptian, and Maghrebi varieties (Habash, 2010; Darwish et al., 2017); (2) *culturally embedded hostility* targeting lineage, tribal affiliation, and sectarian identity with no equivalent in English-derived

Proceedings of the 20th Linguistic Annotation Workshop (LAW XX), pages 47–58

categories (Mubarak et al., 2017; Albadi et al., 2018); (3) *sarcasm-heavy pragmatics* in which conventional praise encodes contempt; and (4) *religious rhetoric* in which intra-sectarian delegitimization is misclassified by generic religion categories. Each property corresponds to a mechanism through which guideline inadequacy becomes annotation error.

Two clarifications are warranted before proceeding. First, while sarcasm, irony, and culturally specific insults exist in many languages and cultures, what matters here is not their existence in isolation but the specific way they interact with Arabic dialect distribution and with the categorical schemata of dominant English language guidelines. Second, dialectal variation is a general challenge for Arabic NLP, but in the context of hate speech annotation it has a specific consequence we make precise below: a single token can be hostile in one dialect and benign in another, so a guideline that does not stratify by dialect produces systematic disagreement that is correlated with annotator background rather than with the text under annotation.

We identify three mechanisms through which annotation guidelines produce systematic errors in this setting:

1. **Cultural misclassification**: culturally specific hostile expressions are absent from annotation categories, leading to inconsistent labeling across annotators who each follow the guidelines as written.

2. **Dialectal ambiguity**: lexical items whose meaning varies across Arabic dialects are assigned labels based on only one of their possible interpretations, producing systematic disagreement correlated with annotator dialect background.

3. **Annotation projection**: annotation schemes designed for English social categories are transferred to Arabic without adaptation, introducing categorical mismatches that no annotator can resolve within the given framework.

Through linguistic case studies with attested Arabic examples, we illustrate each mechanism and show how it generates recurrent annotation errors. We then propose a five-type taxonomy of guideline-induced errors, an explicit mapping from the three mechanisms to the five error types, and a four-stage diagnostic framework.

2 Background

2.1 Annotation Errors and Human Label Variation

Artstein and Poesio (2008) provide a foundational survey of inter-coder agreement measures and the conditions under which each is appropriate. High IAA has been treated as a proxy for annotation quality, but this equivalence has been repeatedly challenged.

Plank et al. (2014) argue that disagreements in NLP annotation are often linguistically meaningful rather than spurious noise. Pavlick and Kwiatkowski (2019) show that disagreement in NLI persists even with additional context, suggesting real linguistic uncertainty rather than annotator error. Nie et al. (2020) collect ChaosNLI and find that genuine human disagreement is widespread and that current models fail to capture the resulting label distributions. Jiang et al. (2023) elicit ecologically valid explanations from annotators to identify the linguistic and pragmatic sources of label variation, and Giulianelli et al. (2023) broaden the picture to NLG, connecting human production variability to aleatoric uncertainty. Taken together, this body of work establishes that human label variation (HLV) is a pervasive and informative signal, not a defect to be eliminated. Swayamdipta et al. (2020) introduce dataset cartography, a model-based approach that reveals many annotated datasets contain a substantial proportion of ambiguous or mislabeled examples.

Most directly relevant to our work, Klie et al. (2023) systematically analyze annotation error detection across eighteen methods and nine datasets, distinguishing between annotation errors, which are objectively incorrect, and disagreements that reflect plausible variation. No current method reliably separates the two phenomena without human adjudication. Building on this, Weber-Genzel et al. (2024) introduce the VariErr NLI dataset and a two-round annotation methodology for separating annotation errors from label variation, demonstrating that even state-of-the-art automatic methods fall short of human performance.

A complementary strand of work examines how the design stance of annotation guidelines shapes data quality. Röttger et al. (2022) distinguish *prescriptive* paradigms, which discourage annotator subjectivity to enable consistent training, from *descriptive* paradigms, which embrace it to model human diversity. Hate speech annotation sits un-

easily between these poles: prescriptive guidelines are necessary for consistency, yet culturally inadequate ones produce systematic error rather than consensus. Uma et al. (2021) survey methods for learning from disagreement, showing that structured disagreement patterns carry diagnostic value rather than being mere noise. Our work applies this insight to guideline design: the *structure* of disagreement localizes the failure mechanism.

Our contribution focuses on the upstream conditions that make certain errors structurally inevitable, prior to any annotator involvement. We position this against HLV in Section 5: the cases we describe are not legitimate variation to be modeled but underspecification or category mismatch to be repaired.

2.2 Hate Speech Annotation

Hate speech annotation is challenging because hate speech is context-dependent, culturally embedded, and subject to definitional disagreement across research communities, legal frameworks, and platform moderation policies (Waseem and Hovy, 2016; Davidson et al., 2017; Vidgen et al., 2019). Waseem and Hovy (2016) produced one of the earliest large-scale hate speech datasets, annotating English tweets within a framework grounded in Western civil rights categories. Davidson et al. (2017) extended this with a three-way distinction between hate speech, offensive language, and neither. Founta et al. (2018) showed that annotator pool composition substantially affects label distributions. Zampieri et al. (2020) organized a multilingual shared task and showed annotation schemes do not transfer reliably across languages. Vidgen et al. (2021) propose adversarial data collection as a remedy for dataset brittleness.

For Arabic specifically, Mubarak et al. (2017) developed an abusive language dataset using ternary classification, Albadi et al. (2018) constructed the first Arabic dataset for religious hate speech, and Mulki et al. (2019) introduced L-HSAB, a Levantine Twitter dataset. All three were developed with limited attention to dialectal variation and culturally specific insult categories, gaps this paper analyzes as sources of systematic guideline-induced error.

2.3 Arabic NLP and Annotation

The challenges of Arabic NLP are well documented. Habash (2010) surveys Arabic morphological complexity, orthographic ambiguity, and the diglossic relationship between MSA and regional dialects. Darwish et al. (2017) demonstrate that Arabic POS tagging requires feature engineering tailored to Arabic-specific morphological phenomena, underscoring the difficulty of directly applying tools and frameworks developed for English. These challenges compound on social media, where non-standard spelling, code-switching, and high proportions of dialectal content create a particularly heterogeneous linguistic environment (Mubarak et al., 2017).

3 Arabic Social Media Discourse: Annotation Challenges

Arabic social media discourse presents four properties that make it particularly vulnerable to guideline-induced annotation errors. None of these properties is unique to Arabic in the strict sense: sarcasm and indirect insult exist in every language community, and many languages exhibit dialectal variation. The point below is more specific. The *combination* of these properties, together with specific categorical mismatches against dominant English language guidelines, produces a configuration of annotation risks for Arabic that has not been adequately documented.

Dialectal variation as an annotation problem. Arabic is a collection of related varieties in a diglossic relationship with MSA (Habash, 2010). Gulf, Levantine, Egyptian, and Maghrebi Arabic differ substantially at the lexical, morphological, and pragmatic levels (Darwish et al., 2017). On social media, users frequently mix dialects within a single post and use dialect as a marker of regional identity. This is a general fact about Arabic NLP, but it has a specific consequence for hate speech annotation: when a token is hostile in one dialect and neutral or even affectionate in another, an annotator's dialect background biases the label they assign even when they apply the guideline faithfully. The result is disagreement that tracks annotators rather than texts, the diagnostic signature of guideline underspecification rather than legitimate label variation.

Culturally embedded hostility. Arabic hostile discourse frequently targets dimensions absent from English-language hate speech frameworks: lineage and ancestral origin (*nasab*, نَسَب *nasab*), tribal or clan affiliation, regional identity, and sectarian belonging (Mubarak et al., 2017; Albadi et al., 2018). The disparity between the categories

of social harm assumed by guidelines and those operative in the target discourse is the root cause of the errors analyzed here. Lineage and tribe based hostility do exist in Western contexts, but they rarely figure in the core categorical schemata of widely used guidelines, and so they become systematically invisible when those schemata are applied to Arabic data.

Sarcasm and pragmatic indirectness. Arabic online discourse makes extensive use of sarcasm, irony, and indirect criticism (Mulki et al., 2019). Hostile force depends on pragmatic inference rather than explicit surface content, so guidelines focused on surface features systematically misclassify sarcastic praise as neutral. While sarcasm is universal, what is specific here is the dense interaction between conventionalized honorific forms (such as *fandī*) and the pragmatic inversion that contexts of online conflict impose on them. Without explicit guidance on how to read these inversions, annotators have no shared procedure for resolving them.

Religious and sectarian rhetoric. Intra-Islamic sectarian hostility, denying the religious legitimacy of a specific sect, is a prevalent form of group-targeted harm in Arabic online discourse. Generic "religion" categories derived from English moderation policies do not distinguish this from anti-religious speech, creating a categorical gap that produces systematic misclassification (Albadi et al., 2018). As we discuss in Section 5, the boundary between sectarian hostility and ordinary theological disagreement is operationalizable through a small set of textual cues even though it is not trivial.

Honor-based hostility: a closer look. Because honor-linked insult patterns are not transparent to annotators outside the relevant cultural context, we expand briefly on the underlying social logic. In many Arabic-speaking communities, the social standing of a male addressee is publicly tied to the perceived conduct of his female relatives, a construct often referred to as *sharaf* (شَرَاف *šarāf*). Hostile speech can therefore attack the addressee indirectly, by asserting or insinuating that a sister, mother, or wife behaves in ways the local norm system codes as dishonorable. The harm is double layered: it stigmatizes the woman through her imputed conduct and damages the addressee through the implication that he cannot uphold his familial honor. From an annotation perspective, surface-level hostility detection misses these expressions because they typically contain no slur and no protected-attribute reference. The hostile force is recoverable only by an annotator who can read the implicit norm violation, and guidelines that do not surface this pattern make consistent annotation impossible regardless of annotator experience.

4 Three Mechanisms of Guideline-Induced Error

We now present each mechanism with illustrative case studies. The examples are drawn from attested patterns in Arabic social media discourse and are representative of instance types found in existing Arabic hate speech datasets (Mubarak et al., 2017; Albadi et al., 2018; Mulki et al., 2019).

4.1 Mechanism 1: Cultural Misclassification

Cultural misclassification occurs when guidelines define hostility exclusively in terms of categories from one cultural context, leaving annotators without adequate categories for culturally specific hostile expressions. Most hate speech schemes organize hostility around race, ethnicity, gender, sexual orientation, religion, and nationality (Davidson et al., 2017; Waseem and Hovy, 2016), reflecting Western civil rights frameworks that do not exhaust the dimensions of hostility in Arabic online discourse.

Case Study 1: Lineage insults. In many Arabic-speaking communities, lineage (*nasab*, نَسَب *nasab*) carries deep social meaning. Attacking a person's family origin by implying they lack honorable ancestry is a well-established form of hostile expression. The following attested type is common in Gulf Arabic social media:

> **Arabic:** يَا قَدِيّ الأَصل *yā qadiyy al-ʔaṣl*
>
> **Gloss:** "O you of ignoble/lowly origin"
>
> **Note:** Attacks the addressee's family lineage and social standing; highly offensive in Gulf and Levantine contexts.

Table 1 shows a representative annotation pattern for this instance type across three annotators working with a standard hate-speech guideline.

The inconsistency in Table 1 is not annotator error. It arises because the guidelines provide no category for ancestry-based hostility. Annotator B treats lineage as a proxy for social group

Annotator	Label	Rationale
A	insult	personal attack on addressee
B	hate speech	targets a social group
C	offensive, not hate	no protected category matched

Table 1: Annotation disagreement for a lineage insult. The inconsistency arises because guidelines provide no category for ancestry-based hostility, not because annotators have erred.

Context	Variety	Pragmatic Force
descriptive statement	MSA	neutral
hostile comment	Gulf colloquial	strong insult
teasing among friends	Egyptian colloquial	affectionate banter
online argument	Levantine colloquial	mild to strong insult

Table 2: Dialectal and contextual variation in the pragmatic force of *ḥayawān* (حَيَوَان *ḥayawān*) across Arabic varieties.

membership; Annotator C, correctly noting no protected category matches as written, assigns a lower-severity label. Both act in conformity with the guidelines. The error is structural.

Case Study 2: Tribal and regional insults. Insults targeting tribal affiliation or regional origin carry significant social force in Gulf and Levantine contexts. Consider the following example type:

Arabic: أَنتَ مِن قَبِيلَتِهِم المُتَخَلِّفَه *ʾanta min qabiylatihim al-mutaẖalifah*

Gloss: "You are from their backward tribe"

Note: Stigmatizes a named or implied tribal group; functions as group-targeted hate speech in social context but falls outside standard annotation categories.

Such expressions imply membership in a stigmatized tribe or region, functioning as group-targeted hate speech. Annotation schemes that omit regional or within-community ethnicity as protected dimensions produce inconsistent labels that annotator training alone cannot resolve without prior guideline revision.

4.2 Mechanism 2: Dialectal Ambiguity

Dialectal ambiguity arises when guidelines assume stable lexical semantics across Arabic varieties. Many lexical items vary substantially in pragmatic force across dialects, so guidelines without dialect-sensitive examples produce annotation errors even when annotators follow instructions correctly.

Case Study 3: The word *ḥayawān*. The Arabic word for "animal" (*ḥayawān*, حَيَوَان *ḥayawān*) illustrates dialectal ambiguity. In MSA and formal contexts it is a neutral descriptor. Its pragmatic force varies considerably across dialectal and social contexts, as shown in Table 2.

Example A is hostile; Example B is not. A guideline that treats *ḥayawān* as inherently offensive will mislabel Example B. One that treats it as context-dependent but provides no dialect-sensitive guidance produces label variation correlated with annotator dialect rather than post content. This is guideline underspecification, not genuine HLV: the two examples are fully resolvable for annotators who share the relevant dialect. The following attested pair illustrates the divergence concretely:

Example A (hostile, Gulf Arabic):
أَنتَ حَيَوَان وَلَا تِستَحِل الرَدّ *ʾanta ḥayawān walā tistaḥil al-radd*
"You are an animal and do not deserve a response."

Example B (affectionate banter, Egyptian Arabic):
إِنتَ حَيَوَان يَ ابن، كِدَ بِتهِبُّني *ʾenta ḥayawān ya ʾbni, kedá bithibbuniy*
"You animal, my son, is that how you show you care?"

Case Study 4: Honorifics used sarcastically. In Levantine and Gulf Arabic, honorific address forms that are conventionally respectful can be deployed sarcastically to convey contempt. The highly subjective nature of Arabic sarcasm annotation, and its dependence on dialect and cultural context, has been empirically documented: Abu Farha and Magdy (2020) show that annotator dialect background substantially shifts sarcasm labels in Arabic social media data. Consider the following example:

Arabic: تَبَرَكَ الَّه عَلَيكَ يَ فَندِي، مَعرِفَتَك بتُحرِف *tabaraka āllah ʿalayka ya fandiy, maʿrifatak btuḥrif*

Literal gloss: "God bless you, sir, your knowledge is enlightening."

Pragmatic force: Contemptuous sarcasm; the honorific *fandī* and the apparent praise are used ironically to ridicule the addressee's ignorance.

Unlike the *ḥayawān* case, the error here is not a matter of which meaning is primary but of whether pragmatic inversion is recognized as a category of hostile expression at all. Guidelines that do not address sarcasm and do not provide dialect-specific examples of this pattern produce systematic misclassifications. We acknowledge that sarcasm cases sit closest to genuine HLV among the patterns we describe: in some posts, sarcastic and sincere readings are both available, and disagreement reflects real interpretive plurality. Our claim is narrower: even setting those genuinely ambiguous cases aside, a substantial residue of disagreement remains that is attributable to the guideline never naming pragmatic inversion as a category to track.

4.3 Mechanism 3: Annotation Projection

Annotation projection refers to transferring annotation schemes developed for one language or cultural context to a substantially different one without adequate adaptation. In Arabic hate speech annotation this most commonly manifests as English-language moderation categories applied to Arabic discourse, introducing categorical mismatches annotators cannot resolve within the scheme.

Case Study 5: Sectarian hostility. Consider the following attested expression type from Arabic social media:

Arabic: هَاؤُلَاء لَا يَنتَمُونَ إِلَى الإِسلَام الحَقِيقِي
hāʾulāʾ lā yantamūna ʾilā ʾl-ʾislām al-ḥaqiyqiy

Gloss: "These people do not belong to true Islam."

Note: Targets members of a specific Islamic sect; a form of religiously motivated group targeting with severe social consequences in many Arabic-speaking contexts.

Many frameworks derived from English moderation policies collapse sectarian speech into a generic "religion" category that does not distinguish anti-religion from intra-religion hostility, or lack any category for intra-religious targeting. The result aligns with the cross-lingual transfer challenge documented by Zampieri et al. (2020): annotation schemes do not transfer reliably across languages.

Case Study 6: Gender and honor-linked hostility. English-language hate speech frameworks typically treat gender-based hostility in terms of misogyny or sexism directed against women as a social group. Arabic online discourse, however, includes forms of gendered insult structured around the concept of *sharaf* (honor, شَرَاف *šarāf*). Consider the following example:

Arabic: أُختُكَ بِتسَهَر بَرَّه بِلَّليل *ʾuḫtuka bitsahar barrah bil-layl*

Gloss: "Your sister stays out late at night."

Note: An honor-based attack on the male addressee through impugning a female relative's behavior; culturally potent as a hostile move in Gulf and Levantine contexts.

This form of hostility is simultaneously gendered and honor-linked in ways that do not map onto the sexism or misogyny categories of English-derived frameworks: harm targets both the woman discussed and the male addressee whose honor is attacked through association. Guidelines that import English-derived gendered hostility categories without adaptation will systematically under-detect this culturally prevalent form of harm.

5 Taxonomy and Diagnostic Framework

Based on the analysis in Section 4, we propose a taxonomy of guideline-induced annotation errors for Arabic hate speech corpora. The five error types in Table 3 are distinguished by their primary cause in guideline design and by the characteristic misannotation pattern each produces.

This taxonomy is explicitly distinct from human label variation in the sense of Weber-Genzel et al. (2024) and the broader HLV literature (Plank et al., 2014; Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Jiang et al., 2023). The errors we describe are not cases of genuine ambiguity: the guideline itself provides inadequate or misleading instruction, making consistent application impossible regardless of annotator competence. Where HLV may be addressed by modeling disagreement (Uma et al., 2021), guideline-induced errors require upstream revision of the annotation framework. Crucially, the *structure* of observed disagreement signals which revision is needed: clustering by annotator dialect indicates dialectal ambiguity; clustering

Error Type	Guideline Cause	Manifestation	Arabic Example
Cultural misclassification	Category set excludes culturally relevant hostility dimensions	Inconsistent labeling of expressions targeting lineage, tribe, or regional identity	يَا قَدِيّ الأَصل *yā-qadiyy al-ʔaṣl*
Dialectal ambiguity	Guidelines assume uniform lexical semantics across varieties	Same token receives different labels from annotators of different dialect backgrounds	حَيَوَان *ḥayawān* as insult vs. banter
Annotation projection	English-derived categories transferred without adaptation	Sectarian speech collapsed into generic religion label; honor-linked gendered insults misclassified	هَاؤُلَاء لَا يَنتَمُونَ *hāʔulāʾ lā yantamūna*; أُختُكَ بِتسَهَر *ʔuḫtuka bitsahar*
Pragmatic misinterpretation	Guidelines do not address sarcasm or indirect hostility	Sarcastic praise annotated as neutral; ironic honorifics labeled as non-offensive	تَبَرَكَ الَّه عَلَيكَ يَ فَندِي *tabaraka āllah ʿalayka ya fandiy*
Boundary underspecification	Annotation span boundaries undefined for multi-unit expressions	Inconsistent annotation of multi-token or idiomatic insult constructions	Multi-word dialect insult idioms

Table 3: Taxonomy of guideline-induced annotation errors in Arabic hate speech corpora. Error types are distinguished by their upstream cause in guideline design rather than by annotator behavior.

by cultural background indicates cultural misclassification; uniform disagreement suggests boundary underspecification.

5.1 Mapping Mechanisms to Error Types

The relationship between the three mechanisms in Section 4 and the five error types in Table 3 is many-to-many rather than one-to-one, made explicit in Table 4. Cultural misclassification as a mechanism produces the cultural misclassification error type directly and contributes to annotation projection errors when an imported scheme imposes the wrong categorical fit. Dialectal ambiguity produces both the dialectal ambiguity error type and a share of pragmatic misinterpretation errors when dialect licenses the pragmatic inversion. Annotation projection feeds into several error types simultaneously. Boundary underspecification cuts across all three mechanisms whenever the guideline does not specify how to delimit multi-unit hostile expressions.

Mechanism	Error types it primarily produces
Cultural misclassification	Cultural misclassification (primary); annotation projection (secondary)
Dialectal ambiguity	Dialectal ambiguity (primary); pragmatic misinterpretation (secondary, via sarcasm)
Annotation projection	Annotation projection (primary); cultural misclassification (secondary); pragmatic misinterpretation (secondary)
All three	Boundary underspecification (whenever multi-token expressions are involved)

Table 4: Mapping from mechanisms (Section 4) to error types (Table 3). The relationship is many-to-many: a single mechanism can drive several error types, and a single error type can arise from more than one mechanism.

5.2 Extended Category Schema for Arabic

To operationalize the cultural misclassification and annotation projection error types, we propose five culturally grounded hostility dimensions that standard English-derived guidelines omit. For each, we provide a working definition and a positive/negative example pair.

Lineage hostility targets a person's ancestral origin (*nasab*, نَسَب *nasab*). *Positive*: explicit imputation of ignoble ancestry (*yā qadīy al-aṣl*). *Negative*: statements about genealogy that do not stigmatize.

Tribal/regional hostility stigmatizes membership in a named or implied tribal, clan, or regional group. *Positive*: attributing backwardness or inferiority to a named tribe or region. *Negative*: neutral identification of regional origin without derogatory framing.

Sectarian hostility targets members of a specific religious sub-community through exclusion or delegitimization. *Positive*: denying the Islamic cre-

dentials of a sectarian group (*hā'ulā' lā yantamūna ilā al-islām al-ḥaqīqī*). *Negative*: neutral theological disagreement without group-targeting framing. The boundary between sectarian hostility and theological disagreement, while non-trivial, admits an operational test based on three textual cues: (i) the predicate targets the *persons* affiliated with the sect rather than a doctrinal proposition; (ii) the framing involves exclusion from a religiously legitimate ingroup (formulations such as "not real Muslims", "outside true Islam"); and (iii) the expression carries an exhortative or stigmatizing illocutionary force rather than a propositional one. Disagreements with a doctrine in the abstract, without these features, fall outside the category.

Honor-linked gendered hostility attacks a male addressee through the implied behavior of a female relative. *Positive*: insinuating that a sister or mother violates honor norms (*ukhtuka bitsahar barra bil-layl*). *Negative*: neutral reference to a female relative's activities without honor-impugning implicature.

Pragmatic inversion applies to sarcasm and ironic honorifics where surface praise encodes contempt. *Positive*: honorific address combined with contextual cues of ridicule (*tabaraka Allāh 'alayka ya fandī*). *Negative*: sincere use of the same honorific. Where expressions encode multiple dimensions simultaneously, annotators should apply multi-labeling. A priority rule is needed only for span-level annotation of the *primary target of harm*: the directly addressed party takes precedence over secondary targets.

5.3 Diagnostic Framework

We propose a four-stage process for identifying guideline-induced errors during corpus construction, designed as an actionable checklist for dataset builders. The stages apply iteratively, with later findings feeding back into earlier revisions.

Stage	Key diagnostic question
1. Cultural audit	Does the guideline cover hostility targeting lineage, tribe, region, or sect? Are community members and cultural linguists consulted?
2. Dialect calibration	Does within-dialect κ substantially exceed overall κ? Are dialect-sensitive examples provided for known ambiguous items (e.g., *ḥayawān*, *majnūn*)? Use CAMeLira (Obeid et al., 2022) for disambiguation; Gumar (Khalifa et al., 2016) for Gulf examples.
3. Projection check	Were categories imported from another language? Do they cover sectarian targeting and honor-linked gendered hostility (Mubarak et al., 2017; Albadi et al., 2018)? Categories with low pilot coverage must be revised.
4. Pragmatic specification	Does the guideline address sarcasm and ironic honorifics (Röttger et al., 2022)? Is a SARCASM-FLAG label with rationale fields required? Persistent post-training disagreement signals a category gap (Swayamdipta et al., 2020; Abu Farha and Magdy, 2020; Darwish et al., 2017).

Table 5: Diagnostic framework for guideline-induced errors. Each stage targets one error type from the taxonomy and provides a concrete decision criterion for guideline revision.

6 Implications

6.1 For Annotation Practice

Annotation guidelines should be developed by teams combining computational linguistics expertise with cultural knowledge of Arabic-speaking communities (Vidgen et al., 2019). The choice between prescriptive and descriptive paradigms (Röttger et al., 2022) should be made explicitly: prescriptive guidelines are appropriate for consistent moderation but must be prescriptive about *Arabic-specific* hostility dimensions. Annotators should be recruited with known dialect backgrounds, since within-dialect agreement substantially exceeding cross-dialect agreement signals a candidate item set for guideline revision rather than adjudication. Following Swayamdipta et al. (2020), pilot studies should diagnose guideline failures, not only estimate IAA: the disagreement correlation structure localizes the mechanism.

6.2 For NLP Systems

Guideline-induced errors propagate directly into trained models. Models built on datasets with systematic cultural misclassifications will under-detect culturally specific hostility; those trained on dialectally ambiguous labels may learn spurious form-to-label correlations that fail across dialect boundaries. Adversarial data collection (Vidgen et al., 2021) addresses symptoms rather than causes: categorical gaps cannot be closed by collecting more examples within the same scheme. The errors we describe are not random noise that larger datasets will average out; they are systematic biases that larger datasets will amplify.

6.3 For Dataset Documentation

Dataset papers should include a *guideline scope statement* specifying which cultural dimensions of hostility the annotation scheme is and is not designed to capture. Users of corpora such as Mubarak et al. (2017), Albadi et al. (2018), and Mulki et al. (2019) need to know that models trained on these datasets may under-detect lineage-based, sectarian, or honor-linked hostility, not because the datasets are flawed in execution but because their guidelines were not scoped to cover these dimensions. The cultural adequacy audit (Stage 1) requires community participation beyond standard crowdsourcing (Vidgen et al., 2019).

7 Related Work

Klie et al. (2023) argue that annotation quality requires both detecting errors in existing datasets and preventing them in new ones; our paper addresses prevention via a diagnostic framework for guideline-induced errors. The prescriptive versus descriptive paradigm distinction (Röttger et al., 2022) is directly relevant: prescriptive guidelines are necessary but produce systematic error when culturally misaligned with the target discourse. Uma et al. (2021) show that disagreement patterns carry diagnostic value; our framework operationalizes this to localize the active error mechanism. The broader HLV literature (Plank et al., 2014; Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Jiang et al., 2023; Giulianelli et al., 2023; Weber-Genzel et al., 2024) provides the conceptual scaffolding within which our work is positioned: it argues for taking disagreement seriously as a signal rather than noise, and we extend that argument by carving out a class of disagreement that signals upstream guideline failure rather than downstream interpretive plurality.

Research on Arabic hate speech has grown substantially. Mubarak et al. (2017), Albadi et al. (2018), and Mulki et al. (2019) established foundational datasets whose annotation guidelines have not been systematically examined for the mechanisms we describe. Haj Ahmed et al. (2025) show that L-HSAB exhibits a strong Lebanese-dialect bias, causing models to generalize poorly across other Levantine varieties: a concrete instance of dialectal ambiguity and annotation projection combining to produce dataset brittleness. The need for culturally grounded categories extends beyond hate speech. Al-Khalifa et al. (2026) introduce ADAB, a large-scale Arabic politeness dataset covering MSA and four dialect groups annotated according to Arabic linguistic traditions and pragmatic theory, demonstrating the viability of Arabic-specific annotation schemata grounded in sociopragmatic norms. Abu Farha and Magdy (2020) demonstrate that annotator dialect background shifts sarcasm labels in Arabic social media, providing direct evidence for the dialectal ambiguity mechanism and supporting our Stage 2 recommendation. Zampieri et al. (2020) document cross-lingual transfer failures, and Founta et al. (2018) show that annotator pool composition substantially affects label distributions; our contribution is demonstrating that disagreement *structure* also localizes the active taxonomy error type.

While our analysis focuses on Arabic, the mechanisms are not in principle Arabic-specific. Any language whose discourse community exhibits a diglossic structure, a dialect continuum, or culturally embedded hostility lacking equivalents in dominant English-derived frameworks is in principle exposed to analogous risks. We do not have the empirical basis here to demonstrate parallel mechanisms in specific other languages, and we therefore present this generalization as a conjecture for future work rather than a demonstrated finding.

8 Conclusion

This paper argues that a substantial class of annotation errors in Arabic hate speech corpora originates from structural weaknesses in guideline design rather than annotator failure. We identify three mechanisms, cultural misclassification, dialectal ambiguity, and annotation projection, and show through six case studies how each produces systematic errors distinct from normal label variation. We propose a five-type taxonomy, an explicit mapping from mechanisms to error types, and a four-stage diagnostic framework instantiated for Arabic with specific lexical items and expression types. The paper reframes annotation quality as a problem of guideline design rather than annotator behavior, intervening before annotation begins by targeting upstream sources of error. For Arabic and other languages whose discursive cultures are poorly represented in existing frameworks, this reframing is crucial for building NLP systems that are both accurate and fair.

Limitations

The case studies presented in this paper are illustrative rather than empirical: they demonstrate the plausibility and linguistic basis of the error mechanisms we identify but do not constitute a quantitative analysis of error rates in specific datasets. We therefore frame the present contribution as theoretical and programmatic rather than empirical. The diagnostic framework in Section 5 has not yet been applied end-to-end to a complete dataset, and the operational thresholds we suggest (such as comparing within-dialect to overall agreement) need calibration through use. Reviewers of the present version raised this point pointedly, and we agree it is the most important next step.

A full empirical evaluation would require access to complete annotation logs with annotator-level data and dialect metadata for existing Arabic hate speech datasets, which is not uniformly available in the current literature. Future work should combine the diagnostic framework proposed here with the annotation error detection methods surveyed in Klie et al. (2023) to produce quantitative estimates of error rates attributable to each mechanism. A controlled pilot study applying Stages 1–4 to a subset of an existing dataset such as L-HSAB (Mulki et al., 2019) would allow before-and-after measurement of IAA and dialect-conditioned agreement, providing concrete evidence for the gains our framework is designed to produce. We regard the present work as providing the theoretical grounding and operational vocabulary that such a study requires.

The taxonomy we propose is based on analysis of Arabic social media discourse and may not be exhaustive. Additional mechanisms of guideline-induced error may exist for Arabic or for other languages with similarly complex sociolinguistic profiles. The generalization to other diglossic or low-resource languages, made cautiously in Section 7, has not been validated empirically and should be read as a conjecture inviting further work rather than as a demonstrated finding.

Finally, although we contrast guideline-induced error with HLV throughout, we acknowledge that in practice the two coexist: a single disagreement event may reflect both genuine interpretive plurality and underspecified guideline categories. Disentangling them at the instance level remains an open problem, and our framework is meant to complement rather than replace methods that target HLV directly.

Ethical Considerations

This paper discusses hate speech and abusive language in Arabic. All examples are drawn from or modeled on attested discourse phenomena in Arabic social media, and none are fabricated for the purpose of demonstrating harmful content. We do not release any datasets or annotation resources as part of this submission.

The diagnostic framework proposed here is intended to improve the quality of hate speech annotation guidelines, enabling more accurate and fair automated detection of harmful content. We encourage researchers who adopt our framework to consult community stakeholders during the cultural adequacy audit stage and to consider the downstream uses of the corpora they construct.

The Arabic linguistic examples in this paper are used analytically and are not intended to propagate or endorse the expressions they exemplify. All examples are presented alongside translations and pragmatic notes to prevent out-of-context misuse.

Acknowledgment

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar Development and Innovation Council (QRDI).

References

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, Marseille, France. European Language Resource Association.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76. ACM.

Hend Al-Khalifa, Nadia Ghezaiel, Maria Bounnit, Hend Hamed Alhazmi, Noof Abdullah Alfear, Reem Fahad Alqifari, Ameera Masoud Almasoud, and Sharefah Ahmed Al-Ghamdi. 2026. ADAB: Arabic dataset for automated politeness benchmarking, a large-scale resource for computational sociopragmatics. In *Proceedings of the Fifteenth biennial Language Resources and Evaluation Conference (LREC*

2026), Palma, Mallorca, Spain. European Language Resources Association. To appear.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, and Mohamed Eldesouki. 2017. Arabic POS tagging: Don't abandon feature engineering just yet. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 130–137. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, pages 512–515.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? Evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Ahmed Haj Ahmed, Rui-Jie Yew, Xerxes Minocher, and Suresh Venkatasubramanian. 2025. Navigating dialectal bias and ethical complexities in Levantine Arabic hate speech detection. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 103–108, Abu Dhabi, UAE. Association for Computational Linguistics.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine de Marneffe. 2023. Ecologically valid explanations for label variation in NLI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10622–10633, Singapore. Association for Computational Linguistics.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of Gulf Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4282–4289.

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation error detection: Analyzing the past and present for a more coherent future. *Computational Linguistics*, 49(1):157–198.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56. Association for Computational Linguistics.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Ossama Obeid, Go Inoue, and Nizar Habash. 2022. CAMeLira: An Arabic multi-dialect morphological disambiguator. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 319–326, Abu Dhabi, UAE. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293. Association for Computational Linguistics.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019.

Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–99. Association for Computational Linguistics.

Bertie Vidgen, Trista Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93. Association for Computational Linguistics.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffComEval). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447. International Committee for Computational Linguistics.