

Toxicity to Harm: Dataset Archaeology of a Decade of Online Abuse Research

Srimonti Dutta

WAI USA Research Labs
srimonti@womeninai.co

Akshata Kishore Moharir

University of Maryland
akshatan@terpmail.umd.edu

Abstract

Online abuse research relies on a substantial measurement infrastructure comprising datasets, shared tasks, diagnostic suites, taxonomies, and evaluation practices that shape which harms become visible, comparable, and citable. We introduce *Dataset Archaeology* as a method for auditing that infrastructure and apply it to the first published decade of Abusive Language Online and Workshop on Online Abuse and Harms (ALW/WOAH) research. Our audit combines a reconciled corpus of 203 archival ALW/WOAH papers from 2017-2025 with a 60-resource registry of online-harm datasets and diagnostic benchmarks. The field has expanded beyond early hate-speech, offensive-language, and toxicity benchmarks toward diagnostic, perspectivist, multilingual, and multimodal resources. At the same time, much of the empirical record continues to be shaped by a narrower set of recurring design choices: short-form text, English, Twitter/X, coarse labels, thin context, and limited documentation of labeling authority. Across the dataset and diagnostic resources, documentation of annotation authority, annotator perspective, and disagreement remains uneven. Dataset influence also changes when mentions are separated from central use. Among 404 detected paper-dataset matches for the top reused resources, only 52 were verified as central experimental or analytical uses. We argue that progress in online-harm NLP should be evaluated as progress in measurement instruments, not only model performance, and we offer a next-decade agenda for contextual, cross-cultural, exposure-aware, and stakeholder-legitimate harm measurement.

1 Introduction

Online-harm NLP has spent years turning social conflict into trainable supervision. Over the past decade, hate speech, offensive language, toxicity, misogyny, racism, counter-speech, implicit hate,

harmful memes, and adversarial abuse have all been operationalized through labels, taxonomies, diagnostic examples, shared tasks, and benchmark splits. These resources have been indispensable for empirical progress and they have also shaped the kinds of harm that online-harm NLP has been able to recognize, measure, and treat as central.

This history motivates attention to both model performance and the forms of online harm that the field’s datasets make detectable. Dataset design determines which platforms enter the evidence base, which identities are represented, whose judgments inform the labels, which context survives preprocessing, whether disagreement is preserved, and which metric later stands in for progress. As datasets are reused, these design choices are inherited by baselines, leaderboards, surveys, safety evaluations, and policy-facing analyses. Benchmarks therefore evaluate models within particular operationalizations of harm and can help establish those operationalizations as recurring conventions.

Existing surveys have mapped hate-speech and abusive-language detection as modeling areas (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). A parallel line of work on dataset documentation has argued that data resources need explicit accounts of provenance, intended use, limitations, social context, and maintenance (Bender and Friedman, 2018; Gebru et al., 2021; Hutchinson et al., 2021; Paullada et al., 2021). We bring these concerns together by treating online-harm datasets as measurement infrastructure. This framing shifts attention from the question of whether models can detect abuse to the prior question of what the field has made detectable in the first place. Two systematic reviews are relevant to the resource-level analysis of this study. Vidgen and Derczynski (Vidgen and Derczynski, 2020) examine abusive-language training data and the consequences of recurring dataset construction choices, while Poletto et al. (Poletto et al., 2021) survey hate-speech resources

across languages, annotation schemes, and availability. Dataset archaeology extends this line of work by linking resource design to a longitudinal venue record and to later patterns of reuse. The analysis follows how collection, labeling authority, context, and documentation enter subsequent research practice, including the distinction between bibliographic reference and central experimental or analytical use.

The Abusive Language Online workshops and the Workshop on Online Abuse and Harms (ALW/WOAH) provide a useful archive for this kind of study. Their specialist orientation makes datasets, taxonomies, diagnostic resources, and shared-task conventions especially visible, making the proceedings a useful lens on how measurement practices are consolidated. This archive may differ from the balance of topics and methods found across broader publication venues. The venue began with ALW and later became the WOA, reflecting a broader movement in NLP from relatively narrow detection tasks toward a more sociotechnical understanding of harm. That shift makes the venue a natural site for asking how online-harm research has changed over time, which resources became central, and which assumptions about harm were stabilized through repeated measurement.

We call our approach *dataset archaeology*. We use the term *dataset archaeology* to extend conventional dataset-survey work by examining the assumptions encoded in available resources. A survey describes available resources and organizes them by task, language, platform, or method. An archaeology asks what assumptions those resources encode: how harm was operationalized, who was authorized to label it, what context was preserved, what disagreement was collapsed, and which datasets later became recurring evaluation resources. Figure 1 summarizes the framework that organizes the audit.

Contributions. This work makes the following contributions: (i) we provide a decade-scale empirical account of ALW/WOAH as a measurement venue, based on 203 included archival papers from 2017–2025, (ii) we develop a resource-level coding scheme and apply it to a 60-resource inventory comprising 59 online-harm datasets and diagnostic benchmarks together with one catalogue retained as a meta-resource, covering harm scope, label design, platform source, language, modality, context, stakeholder representation, documentation, and reuse,

(iii) we show that bibliographic presence and central experimental or analytical use can diverge substantially, with the size of that divergence varying across resources; (iv) we use these results to derive a benchmark-design agenda for online-harm measurement. More broadly, we argue that progress in online-harm NLP should be evaluated not only through model performance, but also through the quality of measurement instruments the field builds for studying harm as a sociotechnical process.

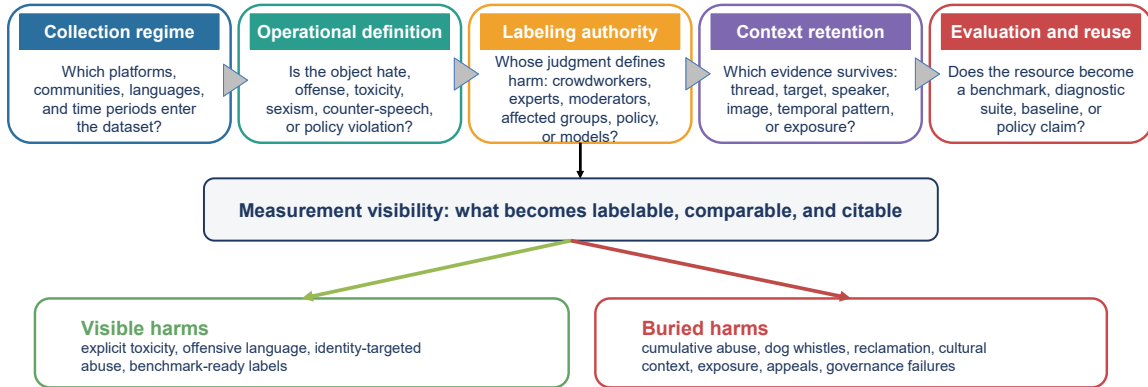
2 Dataset Archaeology

Datasets are often described as inputs to modeling, but that description is too narrow for online harms. Datasets also function as classification systems, governance artifacts, and coordination devices for future research. They can encourage researchers to treat abuse as a property of an isolated text, a relation between speaker and target, a repeated pattern of exposure, a violation of platform policy, a contested social judgment, or a combination of these. Different representations support different research questions by privileging particular units of analysis, forms of evidence, and interpretations of harm.

We use *dataset archaeology* to refer to a structured audit of these measurement choices. The framework examines five sites at which datasets turn online harm into evidence. The first is collection regime: platform APIs, moderation policies, keyword filters, historical access, and language availability shape what can be sampled. The second is operational definition, since hate speech, offensive language, toxicity, harassment, misogyny, counter-speech, and misinformation-as-harm describe related but non-interchangeable constructs. The third is labeling authority, because crowdworkers, researchers, experts, civil-society workers, platform moderators, affected communities, and platform policies bring different forms of knowledge and legitimacy to the annotation process. The fourth is context retention, including whether a dataset preserves threads, targets, speaker history, temporal recurrence, multimodal evidence, or exposure information. The fifth is reuse, through which a resource can become a central benchmark, baseline dataset, diagnostic suite, fairness probe, or simply a cited precedent. Here, measurement visibility includes both explicit labeling and the preservation of relational, cumulative, contextual, indirect, or community-specific evidence needed to recognize a harm. Throughout the paper, *evalua-*

Dataset archaeology audits harm as measurement infrastructure

The framework follows how resources turn social harm into stable, reusable measurement assumptions.



Audit question: What did online-harm NLP learn to measure, and what did its instruments leave outside the frame?

Figure 1: Dataset archaeology audits how online-harm resources turn social harm into measurement assumptions. The framework distinguishes collection regime, operational definition, labeling authority, context retention, and evaluation or reuse. These choices make some harms visible while leaving others outside the measurement frame.

tion practices refers to recurring choices of metrics, benchmark reuse, shared-task protocols, baselines, and diagnostic or fairness evaluation. Table 1 summarizes the purpose of each dimension and the type of distinction it is intended to capture.

This perspective draws on infrastructure studies, where classification systems are understood as durable arrangements that make some forms of work visible while leaving others difficult to see (Star and Ruhleder, 1996; Bowker and Star, 1999). It also builds on critical work in NLP and fairness that cautions against evaluating technical systems apart from the social contexts in which their categories are produced and used (Blodgett et al., 2020; Selbst et al., 2019; Raji et al., 2020). This concern is very relevant to online-abuse research, where alternative annotation and label designs often correspond to substantively different definitions of the target construct. For example, a binary hateful/not-hateful dataset and a perspectivist severity dataset may both be valuable, but they do not support the same theory of harm, the same evaluation practice, or the same downstream moderation decision.

Annotation disagreement provides a concrete example of how measurement design affects the interpretation of a benchmark. Standard supervised-learning pipelines often collapse disagreement through majority vote, which can be reasonable

for narrowly specified tasks but problematic when the underlying judgment depends on lived experience, policy frame, target identity, legal context, or community norm (Waseem, 2016; Aroyo and Welty, 2015; Mostafazadeh Davani et al., 2022). When disagreement is erased, a benchmark can appear more settled than the phenomenon it measures. Dataset archaeology therefore treats disagreement not only as annotation noise, but also as potential evidence about the contested meaning of harm.

3 Evidence Base and Audit Design

The audit has two units of analysis. The paper-level unit is every included archival paper in ALW/WOAH proceedings from 2017 through 2025. The dataset-level unit is an online-harm dataset or diagnostic benchmark that was introduced in the venue, centrally reused by venue papers, or influential in the wider online-harm dataset ecosystem. Figure 2 summarizes the scope. For the claims used in the main analysis, we distinguish automatically extracted fields from targeted human checks. For resource influence, we manually reviewed paper-resource matches for the top automatically detected resources and coded whether each match represented central experimental or analytical use. Candidate pairs were generated by regular-expression matching of resource details in

Audit dimension	Purpose	Illustrative distinction
Collection regime	Characterizes the sampling and access conditions that determine which content can enter a resource.	Keyword- or API-based sampling of public posts versus curated or synthetic examples.
Operational definition	Identifies the harm construct being measured and how it is translated into labels or other targets.	A binary hateful/not-hateful distinction versus a graded severity or target taxonomy.
Labeling authority	Identifies whose judgments establish the annotations and the documented basis of those judgments.	Crowdworker, researcher, expert, community, or policy-based annotation.
Context retention	Records which evidence beyond the focal item remains available for annotation or evaluation.	An isolated post versus a thread with target, speaker, temporal, or multimodal context.
Reuse and evaluation	Characterizes how a resource structures evidence and comparison in later work.	A related-work mention versus use in a main experiment, baseline, diagnostic evaluation, or shared task.

Table 1: The five audit dimensions of dataset archaeology. Each dimension describes a point at which design choices shape what can be represented, interpreted, or compared as evidence of online harm.

Evidence base and audit scope

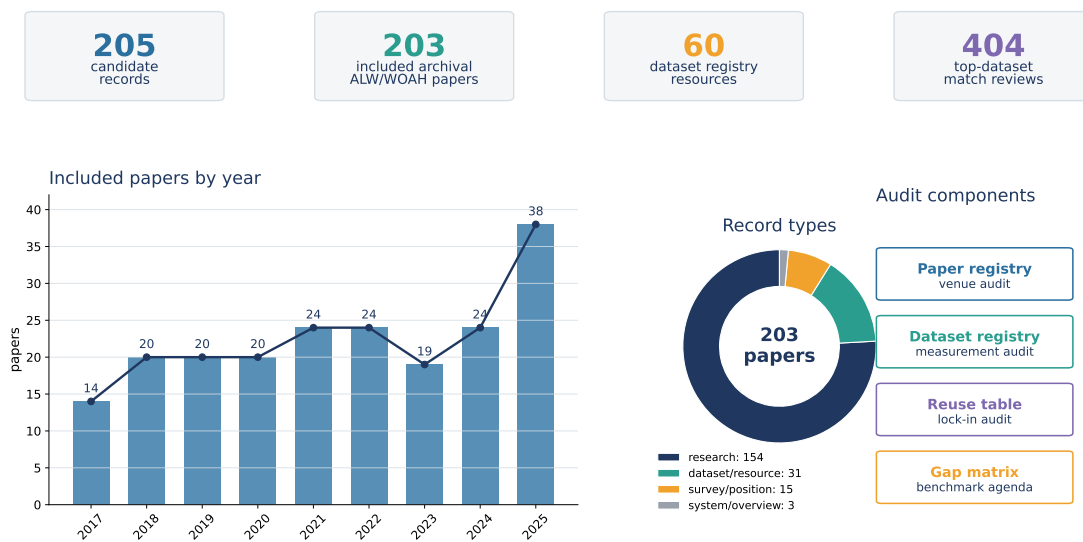


Figure 2: Evidence base and audit scope. The reconciled venue corpus contains 203 included archival ALW/WOAH papers from 2017–2025. The resource registry contains 60 online-harm resources. The reuse analysis reviews 404 candidate paper–resource matches for the ten most frequently detected resources.

paper text; central-use status was assigned only after contextual review, using the operational criteria reported in Appendix A.

Paper corpus. The paper corpus includes 203 archival papers and 2 excluded records. Included records cover research papers, dataset and resource papers, shared-task overviews, system descriptions, surveys, and position papers when they appear in archival proceedings. Excluded records cover front matter, non-paper metadata, duplicate records, and external-pointer stubs.

Resource registry. The registry contains 60 online-harm resources. A resource is included when it supports computational analysis of online

abuse, hate, toxicity, harassment, counter-speech, moderation, misinformation-as-harm, extremism, or closely related harmful online content. The registry records provenance, platform source, language, modality, harm category, label schema, target labels, context availability, annotator type, annotator demographics, disagreement, documentation, ethics, evaluation, and reuse. External resources were selected using explicit eligibility criteria rather than a fixed citation-count or usage threshold. Further operational definitions and details, coding provenance, candidate-matching procedures, and illustrative coding decisions are reported in Appendix A.

Label design	What it measures	Common limitation	n	Representative resources
Binary	Harmful vs. not harmful distinction	Severity, target, disagreement, context, and policy ambiguity	10	HateCheck (Röttger et al., 2021); Multilingual HateCheck (Röttger et al., 2022); Build it Break it Fix it (Dinan et al., 2019).
Continuous	Graded severity or perspectivist measurement	Discrete intervention boundaries and item-level rationales	2	Measuring Hate Speech (Kennedy et al., 2020; Sachdeva et al., 2022); WikiTalk personal attacks (Wulczyn et al., 2017).
Hierarchical	Layered harm and target taxonomy	Out-of-taxonomy harms and contextual harm	9	OLID/OffensEval (Zampieri et al., 2019); OffensEval 2020 (Zampieri et al., 2020); EDOS (Kirk et al., 2023).
Multiclass	Multiple discrete harm categories	Severity within class, spans, and fine-grained intent	4	Davidson et al. (Davidson et al., 2017); Waseem and Hovy (Waseem and Hovy, 2016).
Multilabel	Multiple harm categories per item	Ordinal severity and free-text rationale	1	MLMA Hate Speech (Ousidhoum et al., 2019).
Span/rationale	Localized evidence or rationales associated with harmful content	Social context, target relation, and exposure	2	HateXplain (Mathew et al., 2021); Toxic Spans (Pavlopoulos et al., 2021).

Table 2: Label schemas and their measurement affordances. This is a descriptive summary of the label-schema evidence.

4 What Online-Harm NLP Learned to Measure

We use the paper corpus and dataset registry to examine how online-harm NLP has operationalized harm as an object of measurement. The ALW/WOAH record base and the resources connected to it have been used for this purpose. The analysis focuses on four aspects of the field’s measurement infrastructure: the scope of harms and resources that entered the literature, the label schemas through which harm was made measurable, the visibility of annotation authority and documentation, and the difference between dataset mentions and central dataset use. The analysis indicates substantial expansion in the range of harms represented, alongside continued use of several established benchmark conventions: the field has broadened well beyond early hate-speech and offensive-language classification, but many benchmarks still rely on short-form text, coarse labels, limited context, and uneven documentation.

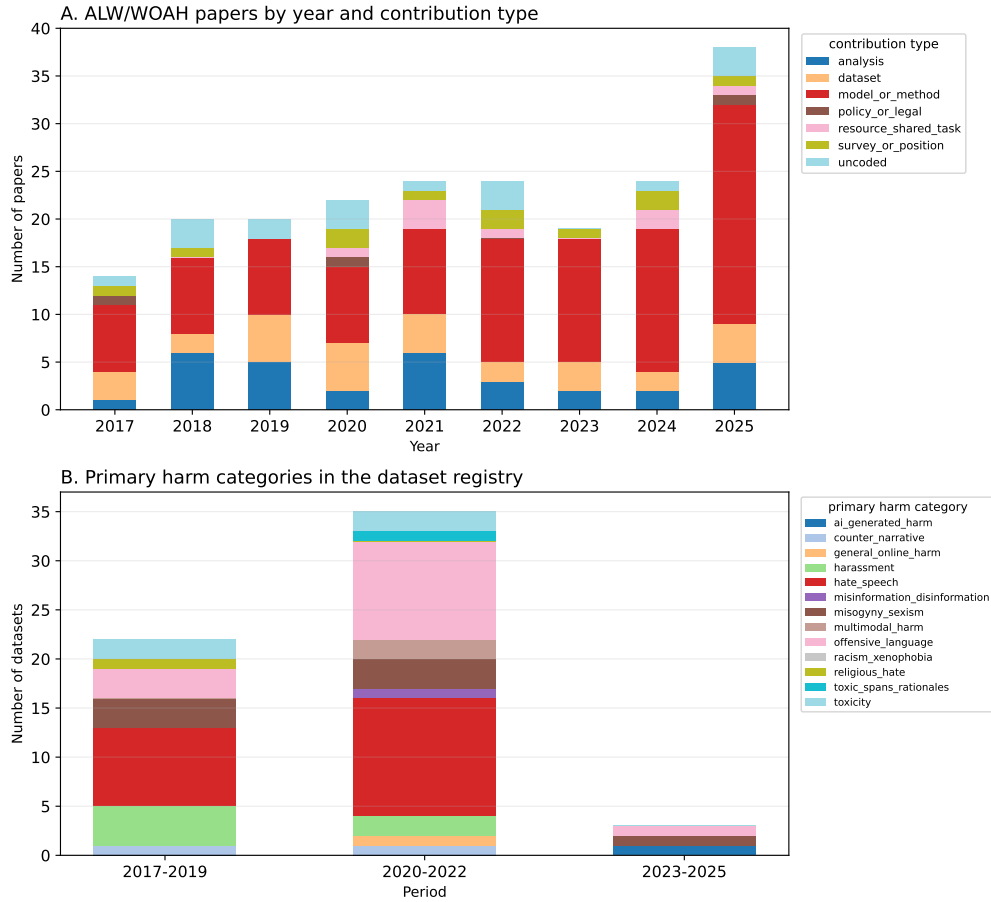
4.1 A Growing Field Built on Familiar Benchmarks

The reconciled ALW/WOAH corpus contains 203 included archival papers from 2017-2025. The yearly record grows from 14 included papers in 2017 to 38 in 2025, with the corpus consisting primarily of research papers, dataset or resource papers, and survey or position papers. The dataset registry contains 60 online-harm resources, including

datasets introduced in or reused by ALW/WOAH papers as well as influential external resources from the broader online-harm ecosystem. Figure 3 summarizes the field’s expansion at two levels: the contribution types appearing in ALW/WOAH papers and the harm categories represented in the dataset registry.

The corpus shows a broadening of the field’s harm vocabulary as represented in ALW/WOAH. Early work was strongly shaped by Twitter/X hate-speech and offensive-language resources, including Waseem and Hovy (Waseem and Hovy, 2016), Davidson et al. (Davidson et al., 2017), and OLID/OffensEval (Zampieri et al., 2019). Later resources extended the measurement landscape toward multilingual offensive-language identification (Zampieri et al., 2020), explainable hate detection (Mathew et al., 2021), counter-narratives (Chung et al., 2019), implicit hate (ElSherief et al., 2021), diagnostic test suites (Röttger et al., 2021, 2022), adversarially generated hate data (Vidgen et al., 2021), machine-generated implicit toxicity (Hartvigsen et al., 2022), and multimodal memes (Kiela et al., 2020). The expansion occurred largely within an enduring measurement paradigm. Many influential resources represent harm as an item-level classification problem over short text. That framing has supported reproducible benchmarks and shared-task evaluation, but it fits some forms of harm better than others and is particularly well suited to harms that can be assessed at the level of individual items. Harassment, abuse, dog whistles,

Field expansion in ALW/WOAH papers and online-harm datasets



Panel A shows the growth and composition of the venue record. Panel B shows how the dataset registry broadened beyond early hate/offensive/toxicity resources.

Figure 3: Field expansion in ALW/WOAH papers and the online-harm resource registry. Panel A shows published ALW/WOAH papers by year and primary contribution type. Panel B shows primary harm categories in the resource registry grouped into three periods.

reclamation, escalation, and exposure-weighted harm often depend on interactional, temporal, or community-specific context that generally requires evidence beyond an isolated post. The historical pattern is thus best understood as expansion around durable measurement conventions, with newer resource designs adding to earlier approaches.

4.2 Label Schemas and the Limits of Benchmark Claims

Benchmark results inherit the assumptions of the label schema. Binary and multiclass labels make classification tractable by representing distinctions such as severity, target, intent, context, and disagreement at a coarser level. More structured designs, including hierarchical, multilabel, span-level, rationale-based, and continuous labels, add useful measurement dimensions but require clearer docu-

mentation of annotation procedures and intended use. Table 2 summarizes this populated subset.

These differences affect the claims that can be drawn from benchmark performance. A binary toxicity dataset may support claims about whether a model distinguishes toxic from non-toxic items under a given annotation protocol. It does not, by itself, support claims about the severity of harm, the legitimacy of a moderation action, the identity of the target, or the effect of repeated exposure. Similarly, a span-level dataset can localize harmful text but may still omit the conversational context needed to interpret why the span is harmful. Rationale-bearing datasets such as HateXplain (Mathew et al., 2021) make part of the evidentiary basis for labels more visible, while perspectivist resources can preserve variation in judgments across annotators. These designs extend the field’s mea-

What the registry makes visible about labeling authority and documentation

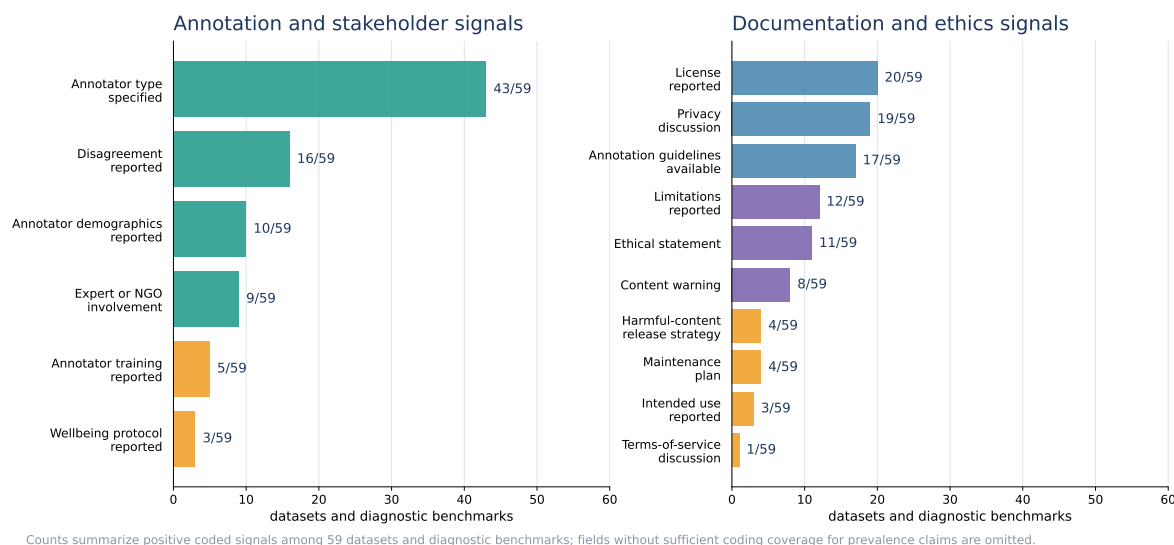


Figure 4: Stakeholder and documentation signals among the 59 datasets and diagnostic benchmarks used for dataset-level analysis. Counts summarize positive signals in the registry; fields without sufficient coding coverage for prevalence claims are omitted.

surement capacity, while also showing that no single label schema can represent the full sociotechnical structure of online harm.

4.3 Labeling Authority and Documentation Gaps

The registry shows that the social basis of annotation is only partly visible and is documented unevenly across resources. Annotator type is recorded for 43 resources and annotator demographics for 10; positive documentation signals for expert or NGO involvement occur in 9 resources and disagreement reporting in 16. Annotator judgments are available in Measuring Hate Speech (Kennedy et al., 2020; Sachdeva et al., 2022) and WikiTalk (Wulczyn et al., 2017). Availability of individual judgments was not systematically checked across the remaining resources, so no overall prevalence estimate is reported for this field. Resources such as SBIC (Sap et al., 2020), ConvAbuse (Cercas Curry et al., 2021), DICES (Aroyo et al., 2023), and AbuseEval (Caselli et al., 2020) illustrate more explicit treatment of annotator perspective, contextual information, or disagreement within this broader record. Figure 4 summarizes these annotation, stakeholder, and documentation signals.

The registry measures visibility of these practices in available documentation, which is distinct from their prevalence in dataset development. These signals capture different aspects of annota-

tion: demographic reporting informs who supplied judgments, procedural documentation records how judgments were elicited and aggregated, and positionality concerns whether the relation between annotator standpoint and target construct is explicitly considered. A dataset may have benefited from community knowledge, expert review, or moderation experience without making those contributions visible in the paper or dataset release. For downstream users, however, undocumented annotation authority is difficult to evaluate or preserve. Later work cannot determine whether the labels reflect crowdworker judgments, researcher decisions, expert guidelines, platform policy, affected-community perspectives, or some combination of these sources. Documentation practices are uneven in other respects as well: within the 59 dataset and diagnostic resources, positive documentation signals are recorded for license information in 20 resources, privacy discussion in 19, annotation guidelines in 17, limitations in 12, and ethical statements in 11. Intended use is recorded for three resources and terms-of-service discussion for one. The significance of these fields varies with resource design. Annotation procedures bear directly on label interpretation, while privacy, licensing, and intended-use documentation are more closely tied to collection, redistribution, and downstream reuse. Privacy documentation is especially relevant for newly collected or user-linked data and less diagnostic for

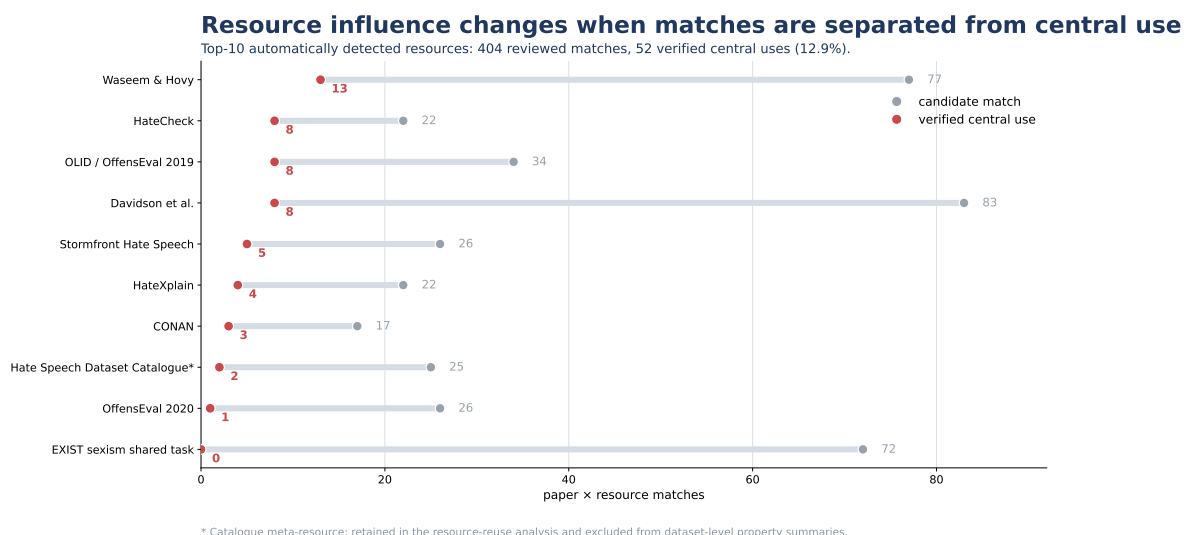


Figure 5: Dataset mentions and verified central use. Across 404 reviewed paper–resource pairs, 52 were verified as central experimental or analytical uses. The Hate Speech Dataset Catalogue is retained here as a catalogue meta-resource and is excluded from dataset-level property summaries

synthetic or previously anonymized material. Likewise, the absence of a separate ethics statement does not by itself indicate that ethical considerations were absent from the study. Appendix B distinguishes the interpretive roles of these documentation fields across different resource settings.

4.4 Dataset Influence Beyond Mentions

Dataset influence is often inferred from citation or mention counts, but the validation results show that this can overstate central benchmark use. For the top automatically detected resources, 404 paper–dataset matches were reviewed. Of these matches, 52 were verified as central experimental or analytical uses. The relationship varies substantially across resources: Waseem and Hovy had 77 candidate matches and 13 central uses, Davidson et al. had 83 and 8, HateCheck had 22 and 8, and EXIST had 72 candidate matches but no central use. For EXIST, 69 matches were false positives and three were related-work references. The reviewed matches thus reveal multiple forms of dataset influence. Some datasets are frequently mentioned because they are part of the field’s background literature. Others are repeatedly used for main experiments, baselines, diagnostic evaluation, fairness analysis, or shared-task participation. Separating these roles provides a more specific account of how a dataset contributes to subsequent work. Related-work mentions indicate bibliographic presence, whereas central use identifies direct dependence in experimental or analytical practice. These

results support coding reuse by function, including related work, main results, baselines, diagnostic evaluation, cross-lingual evaluation, error analysis, and fairness evaluation. This provides a more accurate account of which measurement assumptions are actually amplified through repeated use.

5 A Benchmark Agenda for Online-Harm Measurement

The audit identifies design priorities for the future online-harm benchmarks. Existing resources remain valuable when their scope is clear, but the field needs a wider set of instruments for harms that are contextual, cumulative, cross-cultural, multimodal, exposure-dependent, or tied to moderation practice. Table 3 links the main empirical findings to benchmark directions for the next decade.

These findings motivate two principles for future benchmark design. First, future datasets should document whose judgment is being measured. Crowdsworker judgment, affected-community judgment, expert judgment, moderator judgment, platform-policy judgment, legal judgment, and model judgment should be distinguished explicitly in dataset documentation and annotation design. Second, benchmark design should distinguish classification convenience from harm validity. A benchmark can be easy to score and still fail to measure the harm that motivates the task.

Current default	Audit evidence	Under-measured construct	Benchmark design target
Short-form text classification	Influential early benchmarks are largely organized around item-level text classification.	Interaction, recurrence, escalation, exposure, and cross-platform trajectories.	Thread-, user-, temporal-, and exposure-aware datasets.
Coarse harm labels	Binary, multiclass, and hierarchical schemas dominate populated label-type records.	Severity, ambiguity, intent, target relation, policy uncertainty, and intervention thresholds.	Severity, rationale, target, and appeal-aware label schemas.
Limited disagreement visibility	Disagreement reporting is recorded for a subset of resources; retention of individual judgments was not systematically checked across full dataset-level registry.	Minority views, contested harm perception, and community specific interpretation.	Perspectivist resources that retain item-level judgments and document how disagreement is represented.
Uneven annotator documentation	Annotator type is recorded more consistently than demographic and broader stakeholder information.	Interpretation of whose judgments underlie harm labels.	Annotation protocols with explicit stakeholder roles and visible provenance.
Candidate-match frequency	Candidate matches and central use vary differently across resources.	Direct empirical dependence and the measurement assumptions carried through reuse.	Reuse reporting that distinguishes related work, baselines, main results, diagnostics, and fairness evaluation.

Table 3: Measurement defaults and benchmark design targets. The table connects the audit findings to future dataset designs for contextual, cross-cultural, exposure-aware, and stakeholder-informed online-harm measurement.

6 Discussion and Conclusion

This audit reframes progress in online-harm NLP as a question of measurement infrastructure. Across the first decade of ALW/WOAH, model development occurred alongside the construction of a reusable measurement ecosystem comprising particular harms, platforms, labels, annotators, contexts, and evaluation practices. The historical account in this paper therefore concerns the development of online-harm NLP as represented within the ALW/WOAH communities. Missing information at different stages produces different forms of uncertainty: gaps in construct definitions, annotation procedures, or disagreement limit interpretation of what benchmark labels represent, whereas gaps in licensing, privacy, or intended-use information more directly affect decisions about reuse, transfer, and deployment. The same record shows, however, that expansion is not the same as measurement replacement. Much of the field still makes harm measurable through familiar benchmark forms: short-form text, coarse labels, limited context, and aggregated annotator judgments. These recurring benchmark forms also influence the assumptions under which model performance is interpreted. Through repeated use, they can stabilize assumptions about what harm is, whose judgment counts, which context is necessary, and which forms of evidence are sufficient/can be ignored. The movement from toxicity detection toward broader harm measurement has therefore involved an uneven change in

measurement practice across the field. The reuse analysis distinguishes bibliographic presence from direct empirical dependence. Their relationship varies substantially across resources, as shown by the different candidate-to-central-use patterns in Figure 5. Reporting reuse by function therefore gives a more specific account of how datasets enter later experiments, diagnostics, fairness analyses, and shared-task evaluations. The findings favor explicit statements of dataset validity conditions over a general requirement for broader or more complex resources. In particular, datasets should state their validity conditions: what they measure, the boundaries of that measurement, whose judgment defines the labels, what context was removed, how disagreement should be interpreted, and the deployment settings for which the resource was designed. Without those conditions, benchmark performance can be easily mistaken for harm understanding.

Dataset archaeology offers a way to make these conditions visible at field scale. By reconciling the venue record, coding dataset-level measurement choices, separating central use from mentions, and deriving benchmark-design gaps, it asks what a research community has learned to measure over time. Assessment of the next decade of online-harm NLP should therefore consider improvements on inherited benchmarks together with the development of measurement instruments for contextual, cumulative, cross-cultural, multimodal, exposure-aware, and stakeholder-legitimate harms.

Limitations

The audit is venue-centered. ALW/WOAH is an important archive for online abuse and harms research, but it does not contain all relevant work from ACL, EMNLP, NAACL, ICWSM, FAccT, WWW, CSCW, HCI, policy, or legal venues. The external dataset registry broadens the scope, yet it remains curated rather than exhaustive.

The verification strategy is targeted. Corpus boundary cases, top-dataset reuse relationships, and headline fields for the top reused resources were checked by human review. Many full-registry summaries remain auto-extracted and are labeled as such. The resulting analysis is therefore a broad audit with targeted human validation, while full manual recoding of every dataset field remains a future extension of the present work. Appendix A documents the operational definitions, provenance categories, candidate-matching procedure, and representative coding decisions used in the analysis.

Terminology is unstable across the decade. The meanings and usage of hate speech, offensive language, toxicity, abuse, harassment, misogyny, extremism, and online harm vary across papers and over time. The study normalizes terms for analysis, but normalization may flatten conceptual differences. Thus, the categories should be understood as analytic normalizations rather than a final ontology.

The Harm Measurement Coverage Score is an equally weighted 0–10 descriptive summary of measurement breadth. Narrowly scoped resources may be well suited to narrowly defined tasks, so the score is interpreted alongside resource purpose, scope, year, and intended use. It carries no claim about overall dataset quality. The scoring rubric is reported in Appendix C.

Ethical Considerations

This work audits datasets that contain or describe harmful content. The manuscript reports metadata and documentation patterns while omitting slurs, abusive examples, private messages, and user identifiers. Tables and figures summarize metadata and documentation patterns rather than harmful content. Any public release should include content warnings, avoid redistributing raw abusive content unless permitted by the original license and platform terms, and link to source datasets only when doing so is legally and ethically appropriate.

The paper identifies gaps in documentation,

stakeholder involvement, and reuse practice. We interpret these findings as evidence about changing documentation practices and opportunities for improving future measurement infrastructure. These findings should not be read as personal criticism of dataset creators. Many resources were created before current documentation norms and under severe access constraints. The primary goal of this study is to improve future measurement infrastructure.

Researchers who verify harmful-content datasets may be exposed to distressing material. The recommended audit workflow minimizes direct inspection of abusive examples when metadata and documentation suffice. When inspection is necessary, teams should use opt-out procedures, rotation, debriefing, and local support resources.

References

- Lora Aroyo, Alex Taylor, Mark Díaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. [Dices dataset: Diversity in conversational ai evaluation for safety](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 53330–53342. Curran Associates, Inc.
- Lora Aroyo and Chris Welty. 2015. [Truth is a lie: Crowd truth and the seven myths of human annotation](#). *AI Magazine*, 36(1):15–24.
- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. [Language \(technology\) is power: A critical survey of “bias” in NLP](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.
- Geoffrey C. Bowker and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Inside Technology. MIT Press, Cambridge, MA.
- Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. [I feel offended, don’t be abusive! implicit/explicit messages in offensive and abusive language](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.
- Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. [ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational](#)

- AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. **CONAN - COUNTER NARRATIVES THROUGH NICHESOURCING: A MULTILINGUAL DATASET OF RESPONSES TO FIGHT ONLINE HATE SPEECH**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. **AUTOMATED HATE SPEECH DETECTION AND THE PROBLEM OF OFFENSIVE LANGUAGE**. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–515. Association for the Advancement of Artificial Intelligence.
- Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. **BUILD IT BREAK IT FIX IT FOR DIALOGUE SAFETY: ROBUSTNESS FROM ADVERSARIAL HUMAN ATTACK**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.
- Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. **LATENT HATRED: A BENCHMARK FOR UNDERSTANDING IMPLICIT HATE SPEECH**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Paula Fortuna and Sérgio Nunes. 2018. **A SURVEY ON AUTOMATIC DETECTION OF HATE SPEECH IN TEXT**. *ACM Computing Surveys*, 51(4):1–30.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. **DATASHEETS FOR DATASETS**. *Communications of the ACM*, 64(12):86–92.
- Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. **TOXIGEN: A LARGE-SCALE MACHINE-GENERATED DATASET FOR ADVERSARIAL AND IMPLICIT HATE SPEECH DETECTION**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.
- Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. **TOWARDS ACCOUNTABILITY FOR MACHINE LEARNING DATASETS: PRACTICES FROM SOFTWARE ENGINEERING AND INFRASTRUCTURE**. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, pages 560–575, New York, NY, USA. Association for Computing Machinery.
- Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. **CONSTRUCTING INTERVAL VARIABLES VIA FACETED RASCH MEASUREMENT AND MULTI-TASK DEEP LEARNING: A HATE SPEECH APPLICATION**. *arXiv preprint arXiv:2009.10277*. Revised June 2026.
- Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. **THE HATEFUL MEMES CHALLENGE: DETECTING HATE SPEECH IN MULTIMODAL MEMES**. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, volume 33, pages 2611–2624. Curran Associates, Inc.
- Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. **SEM-EVAL-2023 TASK 10: EXPLAINABLE DETECTION OF ONLINE SEXISM**. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.
- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. **HATEXPLAIN: A BENCHMARK DATASET FOR EXPLAINABLE HATE SPEECH DETECTION**. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875. Association for the Advancement of Artificial Intelligence.
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. **DEALING WITH DISAGREEMENTS: LOOKING BEYOND THE MAJORITY VOTE IN SUBJECTIVE ANNOTATIONS**. *Transactions of the Association for Computational Linguistics*, 10:92–110.
- Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. **MULTILINGUAL AND MULTI-ASPECT HATE SPEECH ANALYSIS**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.
- Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. **DATA AND ITS (DIS)CONTENTS: A SURVEY OF DATASET DEVELOPMENT AND USE IN MACHINE LEARNING RESEARCH**. *Patterns*, 2(11):100336.
- John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. **SEM-EVAL-2021 TASK 5: TOXIC SPANS DETECTION**. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.
- Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. **RESOURCES AND BENCHMARK CORPORA FOR HATE SPEECH DETECTION: A SYSTEMATIC REVIEW**. *Language Resources and Evaluation*, 55(2):477–523.

- Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. [Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing](#). In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20*, pages 33–44, New York, NY, USA. Association for Computing Machinery.
- Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. [Multilingual HateCheck: Functional tests for multilingual hate speech detection models](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. [HateCheck: Functional tests for hate speech detection models](#). 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 41–58, Online. Association for Computational Linguistics.
- Pratik S. Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. [The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism](#). In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.
- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. [Social bias frames: Reasoning about social and power implications of language](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.
- Anna Schmidt and Michael Wiegand. 2017. [A survey on hate speech detection using natural language processing](#). In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.
- Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. [Fairness and abstraction in sociotechnical systems](#). In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, page 59–68, New York, NY, USA. Association for Computing Machinery.
- Susan Leigh Star and Karen Ruhleder. 1996. [Steps toward an ecology of infrastructure: Design and access for large information spaces](#). *Information Systems Research*, 7(1):111–134.
- Bertie Vidgen and Leon Derczynski. 2020. [Directions in abusive language training data, a systematic review: Garbage in, garbage out](#). *PLOS ONE*, 15(12):e0243300.
- Bertie Vidgen, Tristan 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, Online. Association for Computational Linguistics.
- Zeerak Waseem. 2016. [Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter](#). In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. 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, San Diego, California. Association for Computational Linguistics.
- Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. [Ex machina: Personal attacks seen at scale](#). In *Proceedings of the 26th International Conference on World Wide Web, WWW '17*, pages 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. [SemEval-2019 task 6: Identifying and categorizing offensive language in social media \(OffensEval\)](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. 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 \(OffensEval 2020\)](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

A Audit Codebook and Procedures

The audit was organized around a structured set of paper-level, resource-level, and reuse fields. This appendix records the operational definitions used for the principal coded quantities, the provenance of those quantities, and the procedure used to distinguish candidate paper-resource matches from central experimental or analytical use. These definitions make the basis of the reported analyses explicit without depending on a particular software implementation.

A.1 Selected Operational Definitions

Table 4 summarizes the operational definitions for fields that are central to the analyses in the main paper. The table focuses on fields that affect the interpretation of label design, annotation authority, context, and resource reuse.

A.2 Paper-Corpus Reconciliation

The initial archival collection contained 205 candidate records. Manual review covered 64 records, including boundary cases, records with non-standard document types, and a random spot check of ten provisionally standard research papers. All ten spot-checked records were confirmed. The remaining 141 low-priority records were then bulk-confirmed under the same inclusion criteria. This reconciliation yielded 203 included archival papers and two excluded records.

A.3 Candidate Matching and Reuse Verification

Candidate generation used regular-expression matching over the paper text. For the ten resources with the largest automatic match counts, the candidate pairs were subsequently reviewed in context by a single coder. Automatic matches were therefore treated as candidates for review rather than as evidence of dataset use. A detected pair, thus, indicates that a resource name was matched in a paper, whereas its status as central use was determined only during the subsequent human review.

For the ten resources with the largest automatic match counts, all 404 candidate pairs were reviewed in context. Each pair was assigned one of four outcomes: verified central use, related-work-only reference, false positive, or unclear. The review yielded 52 central uses, 282 related-work-only references, 69 false positives, and one unclear case. Thus, 335 of 404 candidates (82.9%) referred to

the intended resource, while 69 of 404 (17.1%) were matching artifacts. This reference-match rate should be distinguished from the 52-of-404 central-use fraction, since related-work references are valid resource matches. Recall is not estimated because paper-resource pairs that did not trigger a pattern were not exhaustively reviewed.

The EXIST acronym illustrates the main source of false positives. Of 72 EXIST candidate pairs, 69 were false positives, three were related-work references, and none was a central use. Most false positives resulted from ordinary occurrences of the word “exist” being matched as the resource acronym.

A candidate pair was coded as a central use when the resource directly supported an experimental or analytical component of the paper, including main results, baseline evaluation, diagnostic evaluation, cross-lingual evaluation, error analysis, fairness or bias analysis, or shared-task participation. Mentions confined to related work or background discussion were coded as non-central, as were textual matches that did not refer substantively to the intended resource. Ambiguous cases for which the available evidence did not support a stable decision were retained as unclear rather than forced into either category. Because the review was conducted by a single coder, no inter-annotator agreement statistic was computed. Consistency within the single-coder design was supported by applying the same operational criteria throughout the review and documenting supporting textual evidence as part of the review process.

A.4 Coding Provenance

The analyses combine automatic extraction with targeted human review. Table 5 records the provenance of the principal quantities so that automatically generated evidence can be distinguished from quantities established through human inspection.

Dataset-level fields were initially populated through automatic extraction from resource papers and associated documentation. Targeted human validation covered ten headline fields for each of the ten resources in the reuse analysis, yielding 100 field-level checks: 83 were confirmed, 16 corrected, and one was not reported. The checked fields covered harm category, platform, languages, number of languages, modality, label structure, annotator type, access status, publication year, and venue. For the label-schema analysis, Table 2 summarizes the 28 of 59 dataset and diagnostic resources with

Field	Operational definition	Primary documentary evidence
Primary harm category	The principal harm construct for which a resource was introduced or used, normalized to the categories employed in the registry.	Resource paper, task description, or accompanying documentation.
Label type	The structure of the target annotation, such as binary, multiclass, hierarchical, multilabel, continuous, or span-level.	Task definition, annotation schema, or resource documentation.
Annotator type	The documented role or identity of the people responsible for assigning annotations.	Methods section, annotation description, or accompanying documentation.
Annotator demographics	Any demographic characteristics of annotators explicitly reported in the available documentation.	Resource paper or accompanying documentation.
Disagreement reported	Whether variation or agreement among annotator judgments is explicitly reported or discussed.	Annotation methods, results, or accompanying documentation.
Context available	Whether the documented resource representation includes contextual information beyond the focal item, such as conversational, target, speaker, temporal, or multimodal context.	Resource description or documented data structure.
Central use	A reviewed paper–resource pair in which the resource directly contributes to an experimental or analytical result, including main evaluation, baseline evaluation, diagnostic or cross-lingual evaluation, error analysis, fairness analysis, or shared-task participation.	Full-text review of the matched paper.
Non-central match	A reviewed paper–resource pair in which the resource occurs as related work, background, descriptive comparison, or a textual matching artifact without direct experimental or analytical use.	Full-text review of the matched paper.

Table 4: Operational definitions used in the audit. Documentary evidence refers to the material from which a field can be coded; the provenance of the quantities reported in this study is summarized separately in Table 5.

populated `label_type` values after targeted corrections.

A.5 External Resource Selection

External resources were eligible for the registry when they satisfied at least one prespecified inclusion criterion. These criteria covered established use or visibility in online-harm research, shared-task use or reuse within ALW/WOAH, inclusion in a major dataset catalogue, a distinctive harm category or annotation design, multilingual or cross-cultural coverage, multimodal or diagnostic design, stakeholder-oriented annotation, and detailed documentation or disagreement modeling. No fixed citation-count threshold was used.

A.6 Illustrative Registry and Reuse Decisions

The following examples show how the coding units used in the analysis are represented. They are included to make the interpretation of the registry and reuse analysis concrete rather than to provide an exhaustive listing of records.

Example registry record. For example, *OffenseEval 2020* is represented in the registry as an offensive-language resource with multilingual coverage and a hierarchical label schema (Zampieri et al., 2020). These fields describe properties of the resource relevant to the audit and are recorded separately from later papers’ uses of that resource. Table 6 illustrates how the reuse coding criteria are applied to central, non-central, and matching-artifact cases.

B Interpreting Documentation Signals

The documentation fields in the registry capture different aspects of resource construction, interpretation, and reuse. Their relevance therefore depends on the type of resource and the decisions that later users need to make. Table 7 summarizes the principal interpretive role of each field.

C Harm Measurement Coverage Score

The Harm Measurement Coverage Score is a descriptive 0–10 summary of measurement breadth. For a resource d , we compute

Analysis component	Provenance	Role in the audit
Venue inclusion and exclusion	Human-reconciled	Establishes the boundary of the archival ALW/WOAH paper corpus.
Full-registry metadata fields	Automatic extraction with targeted human checks	Supports broad resource-level summaries; fields outside the targeted checks are interpreted as coded documentary evidence rather than as a complete manual recoding of every resource.
Candidate paper–resource pairs	Automatically generated	Identifies pairs to be considered for the reuse analysis.
Central versus non-central use	Human-reviewed	Determines whether a candidate pair reflects direct experimental or analytical dependence on the resource.
Reuse function for reviewed pairs	Human-reviewed	Distinguishes roles such as main evaluation, baseline, diagnostic, cross-lingual, error-analysis, fairness analysis, shared-task, or non-central reference.
Headline fields for top reused resources	Targeted human check	Provides additional verification for fields supporting the principal resource-level claims.

Table 5: Provenance of the principal coded quantities used in the audit. The distinction between automatic extraction and targeted human review is maintained throughout interpretation of the results.

Example	Coding outcome	Interpretation
Resource used directly for a reported main or baseline result	Central use	The resource contributes directly to an experimental result and is therefore treated as empirical dependence rather than bibliographic presence alone.
Resource discussed only in related work or background	Non-central use	The resource is part of the paper’s intellectual context but does not directly contribute data or evaluation to the reported analysis.
EXIST candidate matches in the reviewed set	Matching artifact / non-central	The automatic stage produced frequent candidate matches, while manual review found no verified central use among those candidates. This case illustrates why candidate detection and verified use are reported separately.

Table 6: Illustrative application of the reuse coding criteria. The examples distinguish direct experimental or analytical dependence from bibliographic presence and automatic matching artifacts.

$$\text{HMCS}(d) = \sum_{k=1}^{10} I_k(d),$$

where $I_k(d) = 1$ when criterion k is satisfied and 0 otherwise. The ten criteria are: (1) a non-binary label structure; (2) target-group labels; (3) severity, ordinal, or continuous measurement; (4) span, rationale, or explanation labels; (5) context beyond the isolated post, including thread, user, temporal, or external context; (6) disagreement reported or retained; (7) annotator demographics reported; (8) expert, moderator, NGO, or affected-community involvement; (9) multilingual, dialectal, code-switched, or under-resourced-language coverage; and (10) explicit ethics, limitations, or intended-use documentation. Each criterion contributes at most one point.

Higher scores indicate broader measurement coverage and lower scores indicate narrower measurement coverage, which may be appropriate for resources designed for specific tasks. The score is

interpreted alongside resource purpose, scope, year, and intended use.

D Benchmark Gap Matrix

Table 8 expands the shorter benchmark-design table in the main paper. Each row identifies a measurement gap that is difficult to capture with isolated post-level classification alone, explains why that gap is important for online-harm research, and sketches one possible benchmark design. The matrix provides a design map in which individual resources can make focused contributions by addressing particular gaps carefully, documenting their scope, and identifying the forms of harm that remain outside their measurement frame.

E Illustrative Interpretation Cases

The significance of a missing field depends on the resource and on how that resource is used. For example, a missing privacy discussion matters more for newly collected or user-linked platform data than for synthetic material with no identifiable

Documentation signal	Primary interpretive role	Especially informative when
Annotation guidelines	Clarify how the target construct was translated into annotation decisions.	Labels depend on subjective judgments, complex taxonomies, or detailed decision rules.
Annotator demographics	Provide information about the social composition of the people supplying judgments.	Interpretations of harm may vary with lived experience, identity, or social standpoint.
Privacy discussion	Documents consideration of identifiability, sensitivity, and data handling.	Data are newly collected, user-linked, identifiable, or otherwise sensitive.
Ethical statement	Makes ethical considerations and anticipated risks explicit in the documentation.	Collection, annotation, redistribution, or intended application raises substantive ethical concerns.
License information	Clarifies the conditions under which a resource can be accessed, redistributed, or reused.	A dataset is distributed as a reusable benchmark or incorporated into later studies.
Limitations	Identifies known boundaries on interpretation, transfer, or use.	A resource is applied beyond its original task, population, language, or collection setting.
Intended use	States the applications and settings for which the resource was designed.	Benchmark results may be transferred to deployment, moderation, or policy-facing settings.
Terms-of-service discussion	Records platform-related constraints on collection, redistribution, or continued access.	The resource is derived from platform-hosted user content.

Table 7: Interpretive roles of documentation signals used in the dataset audit. The significance of a missing field depends on the resource design, data provenance, and intended use.

Gap	Why is it important?	Possible benchmark design
Contextual abuse	One post may be benign alone and harmful in a thread.	Thread-level annotation with target and speaker context.
Cumulative harassment	Harm can arise from repetition rather than any single message.	Temporal datasets with repeated targeting and exposure.
Dog whistles	Coded hate depends on community knowledge.	Updating lexicons with rationales and community validation.
Identity reclamation	Surface slurs can be harmful or reclaimed.	Speaker and target identity context with policy framing.
Cross-cultural harm	Norms and legal definitions vary across regions.	Multi-region annotation with local guidelines.
Dialect and code-switching	Standard-language models misread dialectal speech.	Dialect-aware and code-switched resources with annotator context.
Multimodal harm	Meaning can be carried by image-text interaction.	Meme, image, video, and audio datasets with aligned rationales.
AI-generated harm	Generative systems create new abuse pathways.	Prompt-output datasets with model provenance and refusal context.
Misinformation-as-harm	False claims can target groups and intensify abuse.	Hybrid disinformation and hate datasets with target labels.
Algorithmic amplification	Ranking can turn borderline content into large-scale harm.	Platform-cooperative datasets with recommendation context.
Exposure-weighted harm	Harm depends on who saw content and how often.	Reach-aware datasets with privacy-preserving exposure measures.
Moderation appeals	False positives and appeals affect legitimacy.	Appeal datasets with policy rationales and outcomes.
Procedural moderation harm	Inconsistent enforcement is itself a harm.	Moderation decision logs with comparable policy categories.
Chilling effects	Over-moderation can suppress vulnerable speech.	Borderline-speech datasets with stakeholder review.

Table 8: Expanded benchmark gap matrix.

users. For a synthetic diagnostic resource, intended use, construct definition, and stated limitations may instead be more important for judging what conclusions the resource can support.

Similar distinctions apply to annotation and context. Missing annotator demographics limits what can be said about whose judgments are represented, while missing annotation guidelines makes it harder to reconstruct how the target construct was translated into labels. Context has a different role again: an isolated-post dataset may be suitable for a narrowly defined classification task but insufficient for studying repeated harassment, conversational escalation, or harms that depend on exposure over time. These cases illustrate why the audit interprets documentation gaps in relation to resource design and intended use rather than treating them as a single measure of dataset quality.