Web(er) of Hate: A Survey on How Hate Speech Is Typed

Luna Wang and **Andrew Caines** and **Alice Hutchings**
Department of Computer Science & Technology
University of Cambridge
Cambridge, UK
{cw829, apc38, ah793}@cam.ac.uk

Abstract

The curation of hate speech datasets involves complex design decisions that balance competing priorities. This paper critically examines these methodological choices in a diverse range of datasets, highlighting common themes and practices, and their implications for dataset reliability. Drawing on Max Weber's notion of ideal types, we argue for a reflexive approach in dataset creation, urging researchers to acknowledge their own value judgments during dataset construction, fostering transparency and methodological rigour.

Warning: This document contains examples of hateful content in Section 6.

1 Introduction

Researchers in computer science, particularly within the NLP community, are increasingly devoting attention to online hate speech. As a deeply social phenomenon, *online* hate speech has been recognised in prior research for its potential to incite and propagate *offline* violence (Lupu et al., 2023). Since Waseem and Hovy (2016), there have been a plethora of hate speech datasets[1] with great diversity in their curation processes despite sharing the overarching goal of advancing state-of-the-art hate speech detection. As noted by previous research, this heterogeneity negatively affects cross-dataset and cross-domain generalisation (Yin and Zubiaga, 2021; Guimarães et al., 2023). At the same time, it has opened up other research directions, such as transfer learning (Ali et al., 2022).

While the differences in datasets are highlighted in past survey studies (Fortuna and Nunes, 2018; Poletto et al., 2021), areas such as design goal and quality assurance are often overlooked. In this paper, we draw on Max Weber's notion of "ideal types" (Weber, 1904, 1930, 1978) (see §2) to highlight that the diversity in hate speech datasets are natural and unavoidable. Instead of pursuing definitional completeness, researchers should adopt a reflexive dataset curation approach. We argue that a fully accurate and comprehensive decomposition of hate speech might not exist. Instead, to progress as a field, the complexities of hate speech should be recognised and the perspectives and assumptions of researchers documented.

We aim to answer the following research question: *After deciding to curate a labelled corpus for hate speech detection, how has past research defined hate speech and how do the design decisions differ?* In doing so, we make the following contributions:

- We apply Weber's ideal types of social action to hate speech datasets, offering a structured framework for understanding socio-political drivers behind hate speech.
- We propose a reflexive approach to dataset curation, encouraging researchers to critically examine and document value judgments and frames of reference to promote transparency.
- We highlight the impact of annotator composition, contrasting smaller, curated annotator pools suited for prescriptive guidelines with more diverse, crowdsourced datasets better aligned with descriptive approaches.
- We critique annotation aggregation practices, advocating alternative ways to capture diverse perspectives and avoid oversimplification.

We provide an overview of Weber's ideal types (§2) and previous surveys (§3). Paper selection is outlined in §4. In §5, we outline key insights and observations. Our discussion (§6) syn-

[1] In this paper, we use the term "hate speech dataset" in its widest sense. We include datasets covering hate speech, abusive language, offensive language, and to a lesser extent harassment and cyberbullying as well as other types of text-based online harms, as described by their corresponding authors.

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 77–103

thesises and interprets our findings. The Appendix includes breakdowns of the datasets analysed.

2 Weber's Ideal Types

The inherent subjectivity and the variability in defining hate speech have been discussed within the NLP community (Fortuna and Nunes, 2018; Vidgen and Derczynski, 2021; Pachinger et al., 2023). This subjectiveness makes hate speech detection as a classification task difficult. In discussing the subjectivity of hate speech detection, Röttger et al. (2022) outline two contrasting paradigms to encourage researchers to either embrace or limit the subjectivity of the task to the fullest extent. Cercas Curry et al. (2024) call for a separation between *-isms* and offence and distinguish individual differences from subjectivity.

Ideal types, conceived by the German sociologist Max Weber, are analytical heuristics that serve to make sense of complex social phenomena. They are not perfectly all-encompassing, nor do they represent the average. Rather, in an observer's attempts to understand phenomena such as capitalism (Weber, 1930) or, more relevant to this discussion, hate speech, these *ideal* constructs are created to "sort out" the underlying complexities. It is therefore inevitable that these constructs depend on the observer's frame of reference, and as a result the observer—whether consciously or unconsciously—articulates certain aspects that they deem worthy while suppressing those of less importance.

Viewed through a Weberian lens, the subjectivity and variation of hate speech datasets are grounded in the frame of reference (cultural norms, historical perspectives, laws, moderation guidelines, and values) that actors (researchers from computer science, linguistics, gender/political/religious studies, criminology or law, annotators, platforms, moderators, speakers, recipients, bystanders, and counter-speech campaigners) choose to adopt and accept. Prescriptive guidelines can limit variation (Röttger et al., 2022), but may still introduce bias through the identity and values of the moderator, speaker, and recipient.

Weber names four ideal types of social action:[2]

Goal-rational (*zweckrational*): motivated by precise and strategic calculation with the aim of achieving some goals.

Value-rational (*wertrational*): motivated by values and beliefs despite their potentially sub-optimal consequences.

Affectual (*affektuell*): driven by emotions.

Traditional (*traditional*): based on established traditions and habits.

In the context of hate speech, **goal-rationality** might see hate speech being used strategically to achieve political or ideological goals. Researchers might be interested in how such discourse polarises public opinions and even radicalises the public to the extremes. From a **value-rational** perspective, hate speech might be expressed in ways that align with the speaker's beliefs about race, gender, or religion. The evaluation of such belief-driven hate speech is heavily dependent on whether the observer (e.g. a researcher, moderator, annotator, or a set of annotation guidelines) shares those values. **Affectual action** hate speech can be an emotional response, such as anger or frustration. This category is relevant when considering hate speech in interpersonal conflicts such as Wikipedia or code repository edit comments. Moderators might struggle with distinguishing these reactionary expressions of emotions from more systematic hate speech. Finally, **traditional** forms of hate speech are embedded in cultural and societal norms and traditions, such as casual misogyny or transphobia in some communities. This, too, requires the observer to be aware of their tradition and how it might affect their judgement of hate.

By operationalising their concept of hate speech, researchers risk missing aspects of discourse that do not fit neatly with their ideal type. For example, anti-Semitic conspiracy theories often do not contain explicit slurs but rely on coded language and misinformation (e.g. accusations of global control) (Rathje, 2021). These types of covert, goal-driven hate have been overlooked by previous ideal types of hate speech. At the same time, however, it is unrealistic and perhaps impossible to create a perfect representation of hate speech. Researchers must rely on using ideal types to study the areas in focus, and any ideal type is an idealised representation, bound to overlook certain aspects.

Actors use frames of reference to construct an ideal type. Goal-rational actions, such as online moderation, may use prescribed guidelines. However, these are not stable, and the terms of reference can change over time and place. Meta and X (formerly Twitter) have changed their policies regarding transphobic hate speech. This highlights

[2] As they are ideal types, they are not mutually exclusive and real world examples often exhibit properties of multiple types at the same time.

the challenge of developing prescriptive guidelines that remain relevant and applicable.

By recognising that any operationalisation of hate speech is an ideal-typical construct, we argue no single decomposition can fully encapsulate the complexity of hate speech. Instead, researchers should explicitly document their perspectives and assumptions, acknowledging the underlying subjectivities in their operationalisation.

3 Related Work

Poletto et al. (2021) provide the most comparable survey of hate speech datasets, reviewing 64 datasets across five dimensions. In contrast, our study doubles the coverage, making it the most comprehensive to date, but adopts a distinct stance on operationalisation. While Poletto et al. (2021) advocate for shared operational frameworks and benchmark resources, we draw on Weberian theory to argue that frameworks and evaluations should be tailored to datasets and models individually in relation to their specific purpose and the curator's ideal-typical operationalisation.

Yu et al. (2024) review 492 datasets, focussing on the targeted identities within hate speech datasets and revealing discrepancies between conceptualised, operationalised, and detected targets, leading to inconsistencies in hate speech classification models. Tonneau et al. (2024) review 75 hate speech datasets across languages and geo-cultural contexts, revealing a diminishing English-language bias but persistent over-representation of countries like the US and UK.

While their work provides valuable insights into identity and geo-cultural representation, our study takes a broader approach by examining the entire dataset curation process, including definitions, intended goals, and design choices. The biases revealed by Yu et al. (2024) and Tonneau et al. (2024) illustrate the gap between curators' ideal types—as conceptualised in their definitions and frameworks—and the realities of their final datasets, reinforcing our argument that dataset validity hinges on alignment with intended objectives rather than definitional completeness.

4 Selection Criteria

The primary source of our datasets is the community-maintained Hate Speech Dataset Catalogue[3] (Vidgen and Derczynski, 2021), which lists 124 research papers and their associated datasets across 25 languages but has limited coverage post-2023. To supplement this, we conducted a Google Scholar search paying particular attention to two venues. Specifically, we conducted two targeted searches and one general search using the following query:

> ("hate" OR "hates" OR "hateful" OR "offensive" OR "offence" OR "offensiveness" OR "harass" OR "harassing" OR "harassment" OR "aggressive" OR "aggressiveness") AND "dataset".

We chose these keywords to broadly cover terms commonly used in existing literature. While we acknowledge scope-specific keywords such as "racism" and "sexism", we did not include those to avoid biasing the search towards specific types of hate.

To target ACL (Association for Computational Linguistics) and ACM (Association for Computing Machinery), we suffix `site:aclanthology.org` and `site:acm.org` to the query respectively. For general search, we append their negative filters to reduce redundancy.

We filter results to studies published from 2023 onward, considering only the first three pages of search results. We only select studies that introduce and describe a new dataset. Non-textual-content-based prediction (e.g. predicting using metadata, Casavantes et al., 2023) are excluded, but re-labelled datasets are included along with their originals.[4] We verify consistency across multiple top-level domains (`.com`, `.co.uk`, `.jp`, and `.hk`). The search is conducted in incognito mode to remove any potential search engine personalisation. We do not conduct a full snowballing process due to its bias toward older studies and limited added value beyond our combined search strategy.

We treat substantially different datasets introduced within the same paper as distinct datasets (e.g. Kumar et al., 2018), as the datasets differ in both data sources and collection methods. In contrast, we regard ETHOS (Mollas et al., 2022) as a single dataset despite its use of two data sources, since other aspects of its creation process remain consistent. In total, we retrieved 135 distinct datasets across 36 languages. Figure 1 shows a breakdown of the number of datasets published in each year, split by source.

[3] hatespeechdata.com

[4] The ACL, ACM, and general searches were conducted on 25 Jan 2025, 3 Feb 2025, and 9 Feb 2025 respectively.

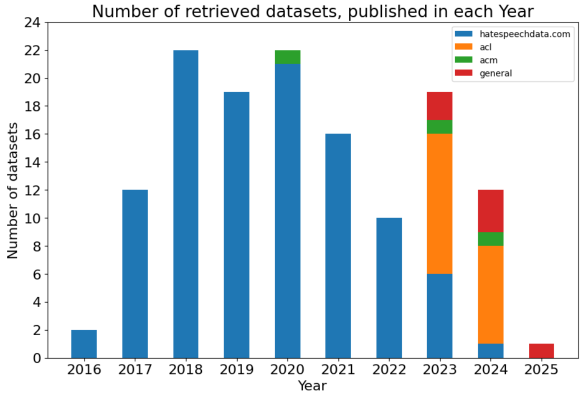

Figure 1: The number of datasets published in each year, split by source of retrieval.

5 Key Insights and Observations

5.1 Frames of Reference

We begin by examining how the authors frame hate speech. Specifically, we look for explicit statements such as "we define hate speech as..." or "hate speech is...". Given the absence, and perhaps impossibility, of a universal definition (Vidgen and Derczynski, 2021; Poletto et al., 2021) and the heterogeneity of the designed tasks, we do not focus on measuring overlap or agreement between definitions. Instead, we identify key areas of coverage and commonly adopted definitions.

Of the 135 datasets, 23 (17%) do not report a definition, and 71 (53%) adopt prior definitions. The remaining 41 (30%) state their own definitions. We analyse the definitions from three overlapping perspectives: 1) categorisation of hate speech into subtypes (e.g., racism, sexism, or categories such as threats and humiliation); 2) specification of the basis for hate (e.g., identities or group affiliations); and 3) referencing of intent (e.g., incite violence, harassment, or insult). Table 1 presents the breakdown of datasets according to these aspects. Among the reported definitions, the basis for hate is most frequently highlighted (60%), followed by subcategorisation (47%) and intent (36%).

5.2 Goals

We examine the designed goals of these datasets, i.e., the research objectives they were designed to achieve. Similar to our analysis of *frames of reference*, we rely on signposting terms such as "aim", "goal", and "to ...". In a number of cases, we infer the aims based on contextual clues without the authors explicitly stating them.

We manually code the stated goals into eight categories: 1) promoting research, new directions, or underrepresented languages ($n = 34$); 2) enabling comparison studies ($n = 3$); 3) supporting automation or model development ($n = 39$); 4) providing finer-grained annotations ($n = 10$); 5) generating insights ($n = 16$); 6) presenting new datasets and resources ($n = 11$); 7) addressing research gaps and challenges ($n = 28$); and 8) benchmarking ($n = 20$). The goals and their associated datasets are listed in Table 2. This shows a considerable proportion of research focusses on automation and model development, exploring new directions in the field, and addressing known challenges.

5.3 Languages

Table 3 shows the distribution of languages. By far, English has received the most attention. The next most frequently studied languages—Italian and German—lag behind by a sizeable margin. There are efforts focusing on multilingual capabilities, as indicated by the mixed-language datasets. Additionally, code-switching has gained traction as a research focus. However, even within code-switched datasets, English remains consistently present, receiving a large portion of attention.

Linguistic variations also play a role in dataset representation. Researchers distinguish between Brazilian Portuguese and European Portuguese, as well as between Mexican Spanish and European Spanish, to account for dialectal differences. Regional and creole languages (Muysken and Smith, 1995), such as Singlish and Hinglish, are included but a strong English basis remains.

Contrary to Tonneau et al. (2024), we did not observe a decline in English datasets' dominance. Instead, compared to non-English datasets, their proportion remains stable in years with more than three retrieved datasets. Possible reasons include different search scopes and methods.

5.4 Data Collection

Datasets are sourced using a variety of methods. Social media platforms dominate, with X/Twitter being the most prevalent data source ($n = 70$). Other platforms include Facebook ($n = 15$), YouTube ($n = 11$), and Reddit ($n = 10$). Instagram ($n = 2$) appears less frequently, likely due to its multimodality. In contrast, traditional online forums are far less represented, with only a handful of datasets sourced from Gab ($n = 4$) and Stormfront ($n = 1$). News website comment sections also serve as a source of online hate

($n = 13$). Additionally, three datasets originate from Wikipedia comments, and two from comments on online code repositories. Beyond data collected "from the wild", some datasets are created "in-house" manually or synthetically ($n = 10$). Other notable sources include language-specific platforms such as Sina Weibo (Jiang et al., 2022) and unconventional sources such as Russian subtitles from *South Park* episodes (Saitov and Derczynski, 2021). Table 4 lists these sources with their respective datasets.

The next step in the dataset creation pipeline is selecting datapoints for annotation. Researchers typically extract a subset of data from a larger corpus. Alternatively, a simpler one-step approach is employed, such as using keyword-based search to directly retrieve relevant instances. We identify three primary techniques for data selection: 1) **Keyword-based sampling** ($n = 73$): searching for relevant content using specific keywords and hashtags. It is the most common method. 2) **Keypage-based sampling** ($n = 26$): focusses on specific recipients or platforms where hate speech is likely to occur. For instance, researchers collect data from key subreddits, Facebook pages, or Twitter accounts by selecting *incoming* comments or tweets. 3) **Keyuser-based sampling** ($n = 25$): unlike keypage-based selection, this technique focusses on the sender rather than the recipient. High-profile users are identified and their *outgoing* comments or tweets are collected.

A subset of datasets ($n = 7$) employ heuristic-based selection methods, applying thresholds to scores generated by external models. These models may be trained on a smaller dataset (Kennedy et al., 2020) or leverage industry solutions such as PerspectiveAPI (e.g., ElSherief et al., 2018; Sarker et al., 2023). Kirk et al. (2023) introduce a unique approach using the score differential between two models as a selection criterion, making it the only dataset to employ a differential-based method.

All but one of the very large datasets ($n = 7$), which contain entries numbering in the millions, do not not use any filtering. Instead, they are comments collected entirely from their respective hosting platforms with their moderation decision. The exception is from Borkan et al. (2019), which is a synthetic dataset.

In terms of languages, geolocation filter ($n = 5$) is commonly used to retrieve language-specific entries, besides data specific sources. Other filtering methods include random sampling (Wulczyn et al., 2017; Moon et al., 2020; Çöltekin, 2020; Kennedy et al., 2022), filtering based on topic (de Pelle and Moreira, 2017; Madhu et al., 2023), and an active-learning-like method (Mollas et al., 2022).

We note many datasets ($n = 47$) use multiple selection methods. When combined, these methods can function either as logical conjunction, i.e. datapoints must satisfy all the requirements to be included, or a logical disjunction, i.e. datapoints are selected if they satisfy at least one requirement.

5.5 Annotation

5.5.1 Task

Hate speech detection can be formalised in various ways as a classification task. These formalisations vary in their granularity, determined by dataset curators' priorities and goals. The simplest and most straightforward approach is binary classification ($n = 34$), where datasets adopt a basic hateful/aggressive/toxic/abusive-or-not framework. While this is easy to implement and operationalise, it lacks nuance, failing to capture meaningful distinctions and subcategories within hate.

Building on the binary classification framework, some datasets ($n = 24$) adopt a multi-class classification approach, where each instance is assigned a single label from multiple (> 2) mutually exclusive categories. This framework provides greater granularity, but it assumes clear-cut distinctions between categories, which may not always be compatible with the ambiguity introduced by edge cases and contexts. For instance, intersectional identities cannot be adequately expressed under this framework. As a result, a model trained by these instances may be biased, as some identities are systematically underrepresented.

Further relaxing the assumption of rigid class boundaries, the multi-label framework ($n = 4$) allows an instance to be assigned multiple applicable labels. In this approach, labels are organised in a flat structure, meaning they are mutually independent and not hierarchically related.

Labels can also be organised hierarchically ($n = 54$), where labels are more structured, and can be tailored towards different levels of granularity. A well-defined taxonomy is essential to this framework. Notably, almost all ($n = 43$) hierarchical datasets rely on an initial binary classification, where the root level question is a binary one. While this approach address the granularity problem, it also inherits the shortcomings of binary classification such as oversimplification. Figure 2 depicts a prototypical hierarchical taxonomy.

We also identify another type of structure, which we refer to as a "parallel" structure ($n = 7$). Unlike hierarchical frameworks that impose a single top-down taxonomy, parallel structures decouple multiple top-level concepts, allowing each to have its own independent internal structure. This approach provides greater flexibility in capturing different aspects of hate speech, as distinct dimensions can be subcategorised separately. For example, Ousidhoum et al. (2019) apply five classification taxonomies in parallel, covering directness, hostility type (including *none*), target, group, and sentiment.

Other types of formalisation include token-level classification (Pamungkas et al., 2020; Pavlopoulos et al., 2021; Saker et al., 2023). This approach offers more interpretability, but puts emphasis on inter-annotator agreement in relation to span boundaries.

Each of these frameworks operationalises different ideal types, emphasising certain aspects of hate while overlooking others. No single framework fully captures the complexity of hate speech. Moreover, even when two datasets adopt the same framework, they may still show inconsistencies due to the differing underlying ideal types of hate, meaning that the apparent similarity in classification structure can be misleading, as differences in these ideal types are not immediately apparent. Thus, a reflexive approach to dataset design, acknowledging and documenting these trade-offs, can lead to more effective and transparent datasets.

5.5.2 Annotators

The majority of the datasets use multiple annotators to label each example, while 13 have only one annotator attending to each example at some stage of annotation. However, in some cases multiple annotators are not feasible, for example when annotators are asked to *construct* sentences (Goldzycher et al., 2024), rather than label them (Table 7).

Subsetting is a popular method to manage multiple annotators, where a (proper) subset of annotators from a pool is assigned to each example ($n = 29$), while others ($n = 47$) assign every annotator to every example. Crowdsourcing ($n = 29$) is a special case of subsetting, where the annotator pool is large and not manually selected.

Among datasets with annotator subsetting, the pool sizes range from as few as three annotators (Pamungkas et al., 2020) to 50 (Romim et al., 2021). Most assign two annotators per instance, though some have up to five. For datasets without subsetting, the highest number of annotators assigned to an example is seven (Pavlopoulos et al., 2021).

Smaller, hand-picked pools can increase annotation consistency, as researchers can enforce a uniform ideal type through additional training and moderation meetings, complementing prescriptive guidelines (Röttger et al., 2022). In contrast, crowdsourcing makes large annotator pools more accessible, potentially increasing demographic diversity, but this is not always guaranteed (Tonneau et al., 2024). A larger pool is better suited for descriptive guidelines, which aim to capture the diversity of human opinions without imposing a predefined ideal type (Röttger et al., 2022). However, under such settings, care must be taken to ensure actual diversity. Transparent reporting of annotator demographics is also vital in datasets with large annotator pools to assess potential biases and ensure a true representation of diverse ideal types.

5.5.3 Annotator Demographics

More than half of the datasets ($n = 78$) do not report annotator demographics. Among those that do, the most commonly mentioned attributes are age ($n = 33$), gender ($n = 33$), and language ($n = 32$). Other reported characteristics include education level ($n = 18$) and location-based information such as nationality ($n = 18$). A smaller number reference sexual orientation ($n = 6$), proxies of socio-economic status (e.g., profession, income) ($n = 10$), political leanings ($n = 3$), or annotators' prior experience with the subject matter, social media, or online abuse ($n = 7$). Table 8 lists a subset of these dimensions.

5.5.4 Disagreements

Most datasets aggregate multiple annotations into a single ground truth label. The utility of this step depends on the dataset's goal. For prescriptive guidelines, where a unified interpretation is intended, assigning a gold label is appropriate. However, for descriptive guidelines that aim to capture the diversity of human judgments, enforcing a single label is counterproductive (Röttger et al., 2022).

To obtain gold labels, many datasets ($n = 48$) use a simple majority rule, while some ($n = 27$) involve additional annotators outside the original pool. Eight datasets resolve disagreements through moderation meetings. Other approaches include positive-class tie-breaking strategy (Gao

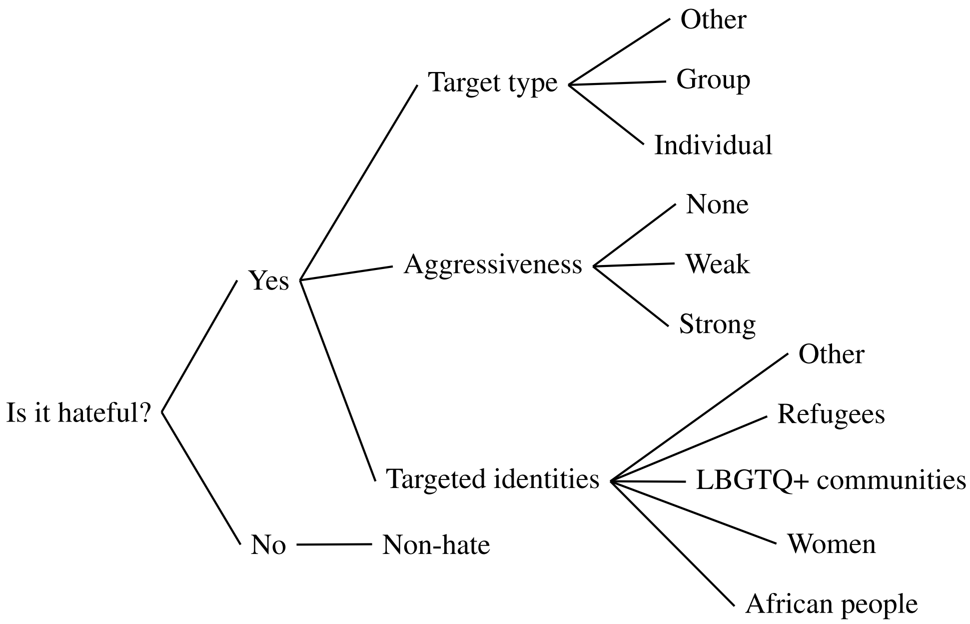

Figure 2: A prototypical hierarchical categorisation of hate speech taxonomy.

and Huang, 2017), and different positive threshold, where the positive label is assigned if positive annotations exceed a threshold (Leite et al., 2020; Assenmacher et al., 2021) (Table 9). Some datasets ($n = 9$) discard instances with disagreement. However, this approach risks losing difficult and ambiguous cases, which can better capture real-world ambiguities, and may reinforce bias.

5.5.5 Quality Assurance

As a final dimension, we examine quality assurance (QA) measures, an often-overlooked aspect in previous surveys. We focus on the steps taken, if any, to ensure dataset quality. Around half of the datasets ($n = 69$) do not report or are unclear about their QA procedures. Of those that do, we observe a relatively even distribution across approaches.

Before annotation, some crowdsourced datasets ($n = 10$) select their annotators based on performance metrics (e.g. approval rate) as well as other data such as geo-location. Some incorporate onboarding training ($n = 13$), which may involve a trial where annotators label a small subset of the data (e.g. Golbeck et al., 2017). Twenty-four datasets employ moderation meetings, though only 10 explicitly mention refining guidelines based on discussions. Annotator tests are also employed by a number of datasets ($n = 12$). These tests can be embedded in the annotation in the form of hidden tests and attention checks, or during onboarding, where annotators that fail a screening test are rejected(Assenmacher et al., 2021; Lee et al., 2024).

Post-annotation QA includes external validation: ten datasets invite external experts to validate a subset of the annotation. Some datasets (e.g. Pavlopoulos et al., 2017; Wiegand et al., 2019) use external annotation and disagreement rates as a proxy for quality. This practice assumes a prescriptive guideline and goal, as high disagreement can still indicate high quality annotation under a descriptive framework (Lee et al., 2024).

5.6 Ethics

Of the papers reviewed, only 14 explicitly revealed they had approval or exemption from an Institutional Review Board (IRB) or ethics committee. A further 27 papers discussed ethical matters, such as anonymisation, but did not reveal if the research had undergone a review process. We note the exclusion of an ethics discussion does not mean the research was not reviewed, or imply that the research was not undertaken ethically. We notice a positive trend, with most of the more recent papers are least partly addressing ethical issues, indicating a growing recognition of the importance of ethics within the research community.

By far the most discussed ethical concern was anonymisation ($n = 21$). One of two approaches are commonly used for anonymisation when releasing datasets, as noted by Cercas Curry et al. (2021). The first approach is to only make an ID (e.g. Tweet ID) available, so that if a user or platform subsequently deletes a post it is no longer available. The second is to make the contents available, but to strip out any identifying information. The possibility that the datasets could be misused was considered in 11 papers, however it was noted that the benefits of the research typically outweighed any potential harm. Some researchers do not make their datasets available due to concerns about misuse (Golbeck et al., 2017; Steffen et al., 2023; Vargas et al., 2024; Wijesiriwardene et al., 2020), while others stipulate restrictions on use (Assenmacher et al., 2021; Fortuna et al., 2019; Lee et al., 2024).

The well-being of annotators, participants, and researchers was discussed in nine papers. Mitigations included allowing annotators to leave at any time (Qian et al., 2019; Vásquez et al., 2023), making mental health support available (Kirk et al., 2022; Lee et al., 2024; Vidgen et al., 2021a), and briefing sessions and regular check-ins (Kirk et al., 2022, 2023). Eight papers also discussed the recruitment of annotators and participants, mainly in relation to compensation. To protect readers and to avoid the perpetuation of harms, authors refrained from providing direct quotes (Cignarella et al., 2024; Kirk et al., 2023; Vidgen and Derczynski, 2021), and provided content warnings (Kirk et al., 2022, 2023).

Only one paper discussed environmental impacts, disclosing the energy sources for their computing clusters (Castillo-lópez et al., 2023). In the future, we anticipate this will become a more prominent consideration, alongside more frequent use of LLMs and awareness of their environmental footprint.

6 Discussions

A Reflexive Approach As hate speech detection inherently involves value judgements, it is crucial for researchers to adopt a reflexive approach throughout the dataset curation process, where the ideal types of hate and curatorial stances are critically examined and reported. In a prescriptive paradigm where disagreements and subjectivity are discouraged, the frame of references of the researchers can still shape their ideal-typical conceptualisation of the categories and definitions. Therefore, researchers must critically examine and document their own value judgements and frame of references as these ultimately shape the annotated datasets and trained models. By making these aspects explicit, researchers can promote transparency and allow for a more nuanced understanding of goal-driven ideal-typical constructs.

Annotator Composition We note the interplay between annotator composition and the author's ideal-typical conceptualisations. Datasets with smaller, hand-picked annotator pools can more easily enforce a uniform ideal type through targeted training and discussions. This approach is more suited for prescriptive guidelines. Conversely, crowdsourced datasets can capture greater diversity, aligning better with descriptive guidelines. However, the persistent underreporting of crowdsourcing annotator demographics presents a challenge in assessing the diversity of captured opinions.

Annotation Aggregation While many datasets rely on majority voting, this method relies on two key assumptions: 1) ground truth is both obtainable and desirable, and 2) annotator consensus reflects this ground truth. Whether these assumptions hold depends on the operational framework. In a descriptive paradigm, aggregating annotations removes the diversity of responses rather than captures it. Additionally, majority voting leaves the underlying sources of disagreement unexamined, further introducing noise. Alternative approaches such as moderation meetings provide a more robust approach for resolving disagreements but are underutilised. Furthermore, datasets that discard instances with disagreement risk removing ambiguous cases, leading to an oversimplification of the task, which may reinforce existing biases.

Application of Ideal Types In this paper, we draw on Weber's notion of ideal types not as categories, but as interpretive lenses reflecting the dataset creators' conceptualisations. In principle, there could be as many ideal types as there are datasets, with each remaining valid within its own context. Rather than attempting to force consensus, the notion of ideal types foregrounds and emphasise the importance of this diversity in curatorial stances.

Furthermore, we suggest the use of Weber's ideal types of social action to interpret hate speech content. While they have not been used as categories to which each dataset is assigned, they can be used as analytical heuristics to interpret the socio-political underpinnings and motivations embedded in these datasets. For instance, PUBFIGS-L (Yuan and Rizoiu, 2025) is a set of manually labelled tweets from 15 American political public figures across the political spectrum. The authors uncover six main themes in hateful and abusive speech: Islam, women, race and ethnicity, immigration and refugees, terrorism and extremism, and American politics (Yuan and Rizoiu, 2025). Through a Weberian lens, such speech can be goal-rational, strategically used to further political agendas, or value-rational, such as religiously motivated hate. Affectual speech aligns with the dataset's category of abuse, distinguishing identity-based hate from emotionally driven personal attacks. The authors also implicitly acknowledge traditional hate speech by noting the presence of covert and implicit hate.

Interpreting using ideal types allows researchers to better understand the heterogeneous curatorial decisions, and better account for the plural underpinnings that motivate hate speech content.

7 Conclusion

Through a Weberian lens, we examine hate speech datasets through Max Weber's ideal types of social action to understand the socio-political underpinnings. We illustrate examples of goal-rational hate, where political figures use hate and abusive language to mobilise the public for political gain, and value-rational hate, where hate speech is driven by ideological beliefs. Moreover, affectual hate can be attacks driven by emotions such as frustration and anger, while traditional hate speech is often normalised and implicit. These ideal types offer a theoretical grounding to the operationalisation of hate speech while acknowledging the diversity of design choices of researchers. Our analysis highlights how dataset construction is shaped by various factors, including the researchers' frame of reference and goal, which in turn influence key design decisions. We advocate for a reflexive approach to dataset construction in which researchers critically examine their own assumptions, operationalisation choices, and the socio-political contexts that shape their work.

Limitations

Our study primarily focusses on publicly available datasets, which may not fully represent the diversity of methodologies used in industry or private research. Second, while we examine key aspects such as frames of reference, goals, languages etc., we do not perform empirical evaluations of annotation quality or dataset performance in downstream tasks. Additionally, our discussion of ideal types and annotation paradigms is necessarily interpretative, and alternative theoretical frameworks could yield different viewpoints.

Acknowledgments

This paper is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 949127).

References

Azalden Alakrot, Liam Murray, and Nikola S. Nikolov. 2018. Dataset construction for the detection of anti-social behaviour in online communication in Arabic. *Procedia Computer Science*, 142:174–181. Arabic Computational Linguistics.

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

Abdullah Albanyan, Ahmed Hassan, and Eduardo Blanco. 2023. Not all counterhate tweets elicit the same replies: A fine-grained analysis. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 71–88, Toronto, Canada. Association for Computational Linguistics.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the Indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238.

Raza Ali, Umar Farooq, Umair Arshad, Waseem Shahzad, and Mirza Omer Beg. 2022. Hate speech detection on Twitter using transfer learning. *Computer Speech & Language*, 74:101365.

Miguel Ángel Álvarez-Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villasenor Pineda, Verónica Reyes-Meza, and Antonio Rico Sulayes. 2018. Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. In *IberEval@ SEPLN*, pages 74–96.

Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz Seiler, Dennis M Riehle, and Heike Trautmann. 2021. RP-Mod & RP-Crowd: Moderator- and crowd-annotated german news comment datasets. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in Hindi. *Preprint*, arXiv:2011.03588.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, Maurizio Tesconi, et al. 2018. Overview of the EVALITA 2018 hate speech detection task. In *Ceur workshop proceedings*, volume 2263, pages 1–9. CEUR.

Uwe Bretschneider and Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media. In *Hawaii International Conference on System Sciences*.

Marco Casavantes, Mario Ezra Aragón, Luis C. González, and Manuel Montes-y Gómez. 2023. Leveraging posts' and authors' metadata to spot several forms of abusive comments in Twitter. *Journal of Intelligent Information Systems*, 61(2):519–539.

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.

Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman, and Malvina Nissim. 2021. DALC: The Dutch abusive language corpus. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 54–66, Online. Association for Computational Linguistics.

Galo Castillo-lópez, Arij Riabi, and Djamé Seddah. 2023. Analyzing zero-shot transfer scenarios across Spanish variants for hate speech detection. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 1–13, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Amanda Cercas Curry, Gavin Abercrombie, and Zeerak Talat. 2024. Subjective isms? On the danger of conflating hate and offence in abusive language detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 275–282, Mexico City, Mexico. 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.

Alessandra Teresa Cignarella, Manuela Sanguinetti, Simona Frenda, Andrea Marra, Cristina Bosco, and Valerio Basile. 2024. QUEEREOTYPES: A multi-source Italian corpus of stereotypes towards LGBTQIA+ community members. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13429–13441, Torino, Italia. ELRA and ICCL.

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France. European Language Resources Association.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. HateMM: A multi-modal dataset for hate video classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):1014–1023.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Rogers de Pelle and Viviane Moreira. 2017. Offensive comments in the Brazilian web: A dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*, pages 510–519, Porto Alegre, RS, Brasil. SBC.

Daryna Dementieva, Valeriia Khylenko, Nikolay Babakov, and Georg Groh. 2024. Toxicity classification in Ukrainian. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 244–255, Mexico City, Mexico. Association for Computational Linguistics.

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. DeTox: A comprehensive dataset for German offensive language and conversation analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to peer hate: Hate speech instigators and their targets. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: A multi-target counter narrative dataset to fight online hate speech. 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 3226–3240, Online. Association for Computational Linguistics.

Paula Ferreira, Nádia Pereira, Hugo Rosa, Sofia Oliveira, Luísa Coheur, Sofia Francisco, Sidclay Souza, Ricardo Ribeiro, João P. Carvalho, Paula Paulino, Isabel Trancoso, and Ana Margarida Veiga-Simão. 2024. Towards cyberbullying detection: Building, benchmarking and longitudinal analysis of aggressiveness and conflicts/attacks datasets from Twitter. *IEEE Transactions on Affective Computing*, pages 1–15.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. *AMI @ EVALITA2020: Automatic Misogyny Identification*, page 21–28. Accademia University Press.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), page 214–228. International World Wide Web Conferences Steering Committee.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Antigoni 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. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 229–233, New York, NY, USA. Association for Computing Machinery.

Janis Goldzycher, Paul Röttger, and Gerold Schneider. 2024. Improving adversarial data collection by supporting annotators: Lessons from GAHD, a German hate speech dataset. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4405–4424, Mexico City, Mexico. Association for Computational Linguistics.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring Hate Speech Detection in Multimodal Publications . In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467, Los Alamitos, CA, USA. IEEE Computer Society.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180, Online. Association for Computational Linguistics.

Samuel Guimarães, Gabriel Kakizaki, Philipe Melo, Márcio Silva, Fabricio Murai, Julio C. S. Reis, and

Fabrício Benevenuto. 2023. Anatomy of hate speech datasets: Composition analysis and cross-dataset classification. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, HT '23, New York, NY, USA. Association for Computing Machinery.

Muhammad Okky Ibrohim and Indra Budi. 2018. A dataset and preliminaries study for abusive language detection in Indonesian social media. *Procedia Computer Science*, 135:222–229. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

Comfort Ilevbare, Jesujoba Alabi, David Ifeoluwa Adelani, Firdous Bakare, Oluwatoyin Abiola, and Oluwaseyi Adeyemo. 2024. EkoHate: Abusive language and hate speech detection for code-switched political discussions on Nigerian Twitter. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 28–37, Mexico City, Mexico. Association for Computational Linguistics.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada. Association for Computational Linguistics.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Cardenas, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and Morteza Dehghani. 2022. Introducing the Gab Hate Corpus: Defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108.

Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted Rasch measurement and multitask deep learning: A hate speech application. *Preprint*, arXiv:2009.10277.

Hannah Kirk, Bertie Vidgen, Paul Rottger, Tristan Thrush, and Scott Hale. 2022. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1352–1368, Seattle, United States. Association for Computational Linguistics.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-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.

Katerina Korre, John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, Ion Androutsopoulos, Lucas Dixon, and Alberto Barrón-cedeño. 2023. Harmful language datasets: An assessment of robustness. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 221–230, Toronto, Canada. Association for Computational Linguistics.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 138–149, Online. Association for Computational Linguistics.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer. 2018. Datasets of Slovene and Croatian moderated news comments. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 124–131, Brussels, Belgium. Association for Computational Linguistics.

Yonatan Lupu, Richard Sear, Nicolas Velásquez, Rhys Leahy, Nicholas Johnson Restrepo, Beth Goldberg, and Neil F. Johnson. 2023. Offline events and online hate. *PLOS ONE*, 18(1):1–14.

Hiren Madhu, Shrey Satapara, Sandip Modha, Thomas Mandl, and Prasenjit Majumder. 2023. Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments. *Expert Systems with Applications*, 215:119342.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018. Did you offend me? Classification of offensive tweets in Hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148, Brussels, Belgium. Association for Computational Linguistics.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: a multilabel hate speech detection dataset. *Complex & Intelligent Systems*.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

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, Vancouver, BC, Canada. Association for Computational Linguistics.

Hala Mulki and Bilal Ghanem. 2021. Let-mi: An Arabic Levantine Twitter dataset for misogynistic language. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 154–163, Kyiv, Ukraine (Virtual). 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, Florence, Italy. Association for Computational Linguistics.

Pieter Muysken and Norval Smith. 1995. The study of pidgin and creole languages. In Jacques Arends, Pieter Muysken, and Norval Smith, editors, *Pidgins and Creoles. An Introduction*, pages 3–14. John Benjamins, Amsterdam, Philadelphia.

Ri Chi Ng, Nirmalendu Prakash, Ming Shan Hee, Kenny Tsu Wei Choo, and Roy Ka-wei Lee. 2024. SGHateCheck: Functional tests for detecting hate speech in low-resource languages of Singapore. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 312–327, Mexico City, Mexico. Association for Computational Linguistics.

Erida Nurce, Jorgel Keci, and Leon Derczynski. 2022. Detecting abusive Albanian. *Preprint*, arXiv:2107.13592.

Anaïs Ollagnier, Elena Cabrio, Serena Villata, and Catherine Blaya. 2022. CyberAgressionAdo-v1: a dataset of annotated online aggressions in French collected through a role-playing game. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 867–875, Marseille, France. European Language Resources Association.

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.

Pia Pachinger, Allan Hanbury, Julia Neidhardt, and Anna Planitzer. 2023. Toward disambiguating the definitions of abusive, offensive, toxic, and uncivil comments. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 107–113, Dubrovnik, Croatia. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Do you really want to hurt me? Predicting abusive swearing in social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6237–6246, Marseille, France. European Language Resources Association.

Hyoungjun Park, Ho Shim, and Kyuhan Lee. 2023. Uncovering the root of hate speech: A dataset for identifying hate instigating speech. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6236–6245, Singapore. Association for Computational Linguistics.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. In *Proceedings of the First Workshop on Abusive Language Online*, pages 25–35, Vancouver, BC, Canada. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-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.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

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.

Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. 2019. Results of the PolEval 2019 shared task 6: first dataset and open shared task for automatic cyberbullying detection in Polish Twitter. In *Proceedings of the PolEval 2019 Workshop*, page 89–110. Polska Akademia Nauk.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. 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 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Md Nishat Raihan, Umma Tanmoy, Anika Binte Islam, Kai North, Tharindu Ranasinghe, Antonios Anastasopoulos, and Marcos Zampieri. 2023. Offensive language identification in transliterated and code-mixed Bangla. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 1–6, Singapore. Association for Computational Linguistics.

Naveen Raman, Minxuan Cao, Yulia Tsvetkov, Christian Kästner, and Bogdan Vasilescu. 2020. Stress and burnout in open source: Toward finding, understanding, and mitigating unhealthy interactions. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results*, ICSE-NIER '20, page 57–60, New York, NY, USA. Association for Computing Machinery.

Jan Rathje. 2021. "Money Rules the World, but Who Rules the Money?" Antisemitism in post-Holocaust conspiracy ideologies. In Armin Lange et al., editor, *Confronting Antisemitism in Modern Media, the Legal and Political Worlds*, pages 45–68. Walter de Gruyter GmbH & Co KG, Berlin.

Akash Rawat, Nazia Nafis, Dnyaneshwar Bhadane, Diptesh Kanojia, and Rudra Murthy. 2023. Modelling political aggression on social media platforms. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 497–510, Toronto, Canada. Association for Computational Linguistics.

Mohammadreza Rezvan, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L. Shalin, and Amit Sheth. 2018. A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, page 33–36, New York, NY, USA. Association for Computing Machinery.

Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in Roman Urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2512–2522, Online. Association for Computational Linguistics.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md. Saiful Islam. 2021. Hate speech detection in the Bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence*, pages 457–468, Singapore. Springer Singapore.

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.

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.

Björn Roß, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the European refugee crisis.

Ramsha Saeed, Hammad Afzal, Sadaf Abdul Rauf, and Naima Iltaf. 2023. Detection of offensive language and its severity for low resource language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(6).

Kamil Saitov and Leon Derczynski. 2021. Abusive language recognition in Russian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 20–25, Kiyv, Ukraine. Association for Computational Linguistics.

Jaydeb Saker, Sayma Sultana, Steven R. Wilson, and Amiangshu Bosu. 2023. ToxiSpanSE: An explainable toxicity detection in code review comments. *Preprint*, arXiv:2307.03386.

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard Jansen. 2018. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. "Call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):573–584.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jaydeb Sarker, Asif Kamal Turzo, Ming Dong, and Amiangshu Bosu. 2023. Automated identification of toxic code reviews using ToxiCR. *ACM Trans. Softw. Eng. Methodol.*, 32(5).

Jaehyung Seo, Jaewook Lee, Chanjun Park, SeongTae Hong, Seungjun Lee, and Heuiseok Lim. 2024. Ko-CommonGEN v2: A benchmark for navigating Korean commonsense reasoning challenges in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2390–2415, Bangkok, Thailand. Association for Computational Linguistics.

Ravi Shekhar, Vanja Mladen Karan, and Matthew Purver. 2022. CoRAL: A context-aware Croatian abusive language dataset. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 217–225, Online only. Association for Computational Linguistics.

Ravi Shekhar, Marko Pranjić, Senja Pollak, Andraž Pelicon, and Matthew Purver. 2020. Automating news comment moderation with limited resources: Benchmarking in Croatian and Estonian. *Journal for Language Technology and Computational Linguistics*, 34(1):49–79.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France. European Language Resources Association.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. MIMIC: Misogyny identification in multimodal internet content in Hindi-English code-mixed language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a WhatsApp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59, Brussels, Belgium. Association for Computational Linguistics.

K. Sreelakshmi, B. Premjith, Bharathi Raja Chakravarthi, and K. P. Soman. 2024. Detection of hate speech and offensive language codemix text in Dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:20064–20090.

Elisabeth Steffen, Helena Mihaljevic, Milena Pustet, Nyco Bischoff, Maria do Mar Castro Varela, Yener Bayramoglu, and Bahar Oghalai. 2023. Codes, patterns and shapes of contemporary online antisemitism and conspiracy narratives – an annotation guide and labeled German-language dataset in the context of COVID-19. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):1082–1092.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott Hale, and Paul Röttger. 2024. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 283–311, Mexico City, Mexico. Association for Computational Linguistics.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Douglas Trajano, Rafael H. Bordini, and Renata Vieira. 2024. OLID-BR: offensive language identification dataset for Brazilian Portuguese. *Language Resources and Evaluation*, 58(4):1263–1289.

Francielle Vargas, Samuel Guimarães, Shamsuddeen Hassan Muhammad, Diego Alves, Ibrahim Said Ahmad, Idris Abdulmumin, Diallo Mohamed, Thiago Pardo, and Fabrício Benevenuto. 2024. HausaHate: An expert annotated corpus for Hausa hate speech detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 52–58, Mexico City, Mexico. Association for Computational Linguistics.

Juan Vásquez, Scott Andersen, Gemma Bel-enguix, Helena Gómez-adorno, and Sergio-luis Ojeda-trueba. 2023. HOMO-MEX: A Mexican Spanish annotated corpus for LGBT+phobia detection on Twitter. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Bertie Vidgen and Leon Derczynski. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32.

Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021a. Introducing CAD: The contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021b. 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.

Max Weber. 1904. Objectivity of social science and social policy. In *Methodology of Social Sciences*. Free Press.

Max Weber. 1930. *The Protestant Ethic and the Spirit of Capitalism*. Routledge, London. Originally published in German in 1905.

Max Weber. 1978. *Economy and Society: An Outline of Interpretive Sociology*. University of California Press, Berkeley. Originally published in German in 1922.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2019. Overview of the GermEval 2018 shared task on the identification of offensive language. In Josef Ruppenhofer, Melanie Siegel, and Michael Wiegand, editors, *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria - September 21, 2018*, pages 1 – 10.

Thilini Wijesiriwardene, Hale Inan, Ugur Kursuncu, Manas Gaur, Valerie L. Shalin, Krishnaprasad Thirunarayan, Amit Sheth, and I. Budak Arpinar. 2020. ALONE: A dataset for toxic behavior among adolescents on Twitter. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings*, page 427–439, Berlin, Heidelberg. Springer-Verlag.

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, page 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Zehui Yu, Indira Sen, Dennis Assenmacher, Mattia Samory, Leon Fröhling, Christina Dahn, Debora Nozza, and Claudia Wagner. 2024. The unseen targets of hate: A systematic review of hateful communication datasets. *Social Science Computer Review*, page 08944393241258771.

Lanqin Yuan and Marian-Andrei Rizoiu. 2025. Generalizing hate speech detection using multi-task learning: A case study of political public figures. *Computer Speech & Language*, 89:101690.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. 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 3181–3197, Online. Association for Computational Linguistics.

Nadezhda Zueva, Madina Kabirova, and Pavel Kalaidin. 2020. Reducing unintended identity bias in Russian hate speech detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 65–69, Online. Association for Computational Linguistics.

A Appendix

A.1 Breakdowns of Reviewed Datasets

Datasets	Subcategories	Basis	Intent	Total
Jha and Mamidi (2017); Salminen et al. (2018); Wiegand et al. (2019); Sprugnoli et al. (2018); Ousidhoum et al. (2019); Borkan et al. (2019); Shekhar et al. (2020); Sigurbergsson and Derczynski (2020); Caselli et al. (2020); Pavlopoulos et al. (2020); Albanyan et al. (2023); Korre et al. (2023); Seo et al. (2024); Ng et al. (2024)	○	○	○	17
Golbeck et al. (2017); Ljubešić et al. (2018); Zampieri et al. (2019); Shekhar et al. (2020); Pitenis et al. (2020); Leite et al. (2020); Saitov and Derczynski (2021); Nurce et al. (2022); Shekhar et al. (2022); Saker et al. (2023); Sarker et al. (2023); Raihan et al. (2023)	●	○	○	13
Roß et al. (2016); de Pelle and Moreira (2017); Fersini et al. (2018); ElSherief et al. (2018); Chung et al. (2019); Qian et al. (2019); Basile et al. (2019); Ibrohim and Budi (2019); Kennedy et al. (2020); Çöltekin (2020); Vidgen et al. (2021b); Grimminger and Klinger (2021); Röttger et al. (2021); Mollas et al. (2022); Ollagnier et al. (2022); Trajano et al. (2024); Kirk et al. (2023); Steffen et al. (2023); Goldzycher et al. (2024)	○	●	○	27
Bretschneider and Peters (2017); Álvarez-Carmona et al. (2018); Suryawanshi et al. (2020); Wijesiriwardene et al. (2020); Kurrek et al. (2020); Caselli et al. (2021); Kennedy et al. (2022); Park et al. (2023); Rawat et al. (2023)	○	○	●	11
Rezvan et al. (2018); Samory et al. (2021)	●	●	○	2
Waseem and Hovy (2016); Waseem (2016); Mubarak et al. (2017)	●	○	●	3
Gao and Huang (2017); Alfina et al. (2017); de Gibert et al. (2018); Mathur et al. (2018); Ptaszynski et al. (2019); Fortuna et al. (2019); Gomez et al. (2020); Romim et al. (2021); Toraman et al. (2022); Kirk et al. (2022); Demus et al. (2022); Castillo-lópez et al. (2023); Saeed et al. (2023); Das et al. (2023)	○	●	●	16
Albadi et al. (2018); Founta et al. (2018); Sanguinetti et al. (2018); Bosco et al. (2018); Mulki et al. (2019); Mandl et al. (2019); Pamungkas et al. (2020); Vidgen et al. (2020); Rizwan et al. (2020); Bhardwaj et al. (2020); Moon et al. (2020); Fersini et al. (2020); Mulki and Ghanem (2021); Vidgen et al. (2021a); Assenmacher et al. (2021); Jiang et al. (2022); Ilevbare et al. (2024); Singh et al. (2024); Yuan and Rizoiu (2025)	●	●	●	22
Mubarak et al. (2017); Wulczyn et al. (2017); Pavlopoulos et al. (2017); Alakrot et al. (2018); Kumar et al. (2018); Bohra et al. (2018); Ibrohim and Budi (2018); Zueva et al. (2020); Raman et al. (2020); Zeinert et al. (2021); Cercas Curry et al. (2021); Pavlopoulos et al. (2021); Fanton et al. (2021); Mathew et al. (2021); Vásquez et al. (2023); Madhu et al. (2023); Cignarella et al. (2024); Vargas et al. (2024); Dementieva et al. (2024); Ferreira et al. (2024); Lee et al. (2024); Sreelakshmi et al. (2024)		*not reported*		24

Table 1: How the definitions are constructed in each dataset. ○: not present, ●: present. Note that one paper may introduce multiple datasets. The number of references and the number of datasets are not necessarily equal.

Coded goals	Datasets	Count
Promoting research, new directions, or underrepresented languages	Waseem and Hovy (2016); de Pelle and Moreira (2017); Wiegand et al. (2019); Kumar et al. (2018); Bohra et al. (2018); Bosco et al. (2018); Álvarez-Carmona et al. (2018); Mandl et al. (2019); Ptaszynski et al. (2019); Fortuna et al. (2019); Kennedy et al. (2020); Gomez et al. (2020); Moon et al. (2020); Fersini et al. (2020); Leite et al. (2020); Çöltekin (2020); Rizwan et al. (2020); Raman et al. (2020); Saitov and Derczynski (2021); Trajano et al. (2024); Rawat et al. (2023); Vásquez et al. (2023); Steffen et al. (2023); Raihan et al. (2023); Saeed et al. (2023); Ilevbare et al. (2024); Vargas et al. (2024); Dementieva et al. (2024)	34
Enabling comparison studies	Waseem (2016); Basile et al. (2019)	3
Supporting automation or model development	Mubarak et al. (2017); Golbeck et al. (2017); Wulczyn et al. (2017); Pavlopoulos et al. (2017); Alfina et al. (2017); de Pelle and Moreira (2017); Alakrot et al. (2018); Sanguinetti et al. (2018); Qian et al. (2019); Shekhar et al. (2020); Sigurbergsson and Derczynski (2020); Vidgen et al. (2020); Pavlopoulos et al. (2020); Zeinert et al. (2021); Samory et al. (2021); Pavlopoulos et al. (2021); Vidgen et al. (2021a); Mollas et al. (2022); Nurce et al. (2022); Kirk et al. (2022); Saker et al. (2023); Sarker et al. (2023); Trajano et al. (2024); Park et al. (2023); Kirk et al. (2023); Saeed et al. (2023); Das et al. (2023); Cignarella et al. (2024); Yuan and Rizoiu (2025)	39
Providing finer-grained annotations	Davidson et al. (2017); Fersini et al. (2018); Founta et al. (2018); Zampieri et al. (2019); Vidgen et al. (2021a); Assenmacher et al. (2021); Shekhar et al. (2022); Kennedy et al. (2022); Demus et al. (2022)	10
Generating insights	Golbeck et al. (2017); Roß et al. (2016); ElSherief et al. (2018); Salminen et al. (2018); Sprugnoli et al. (2018); Ptaszynski et al. (2019); Pamungkas et al. (2020); Pavlopoulos et al. (2020); Cercas Curry et al. (2021); Grimminger and Klinger (2021); Assenmacher et al. (2021); Jiang et al. (2022); Albanyan et al. (2023); Madhu et al. (2023); Cignarella et al. (2024)	16
Presenting new datasets and resources	Rezvan et al. (2018); Chung et al. (2019); Pitenis et al. (2020); Bhardwaj et al. (2020); Moon et al. (2020); Romim et al. (2021); Caselli et al. (2021); Grimminger and Klinger (2021); Fanton et al. (2021)	11
Addressing research gaps and challenges	Gao and Huang (2017); Jha and Mamidi (2017); Albadi et al. (2018); Ljubešić et al. (2018); de Gibert et al. (2018); Mathur et al. (2018); Ibrohim and Budi (2018); Borkan et al. (2019); Ibrohim and Budi (2019); Caselli et al. (2020); Suryawanshi et al. (2020); Fersini et al. (2020); Zueva et al. (2020); Vidgen et al. (2021b); Fanton et al. (2021); Kennedy et al. (2022); Ollagnier et al. (2022); Kirk et al. (2023); Das et al. (2023); Madhu et al. (2023); Goldzycher et al. (2024); Ng et al. (2024); Singh et al. (2024); Ferreira et al. (2024); Lee et al. (2024); Yuan and Rizoiu (2025)	28
Benchmarking	Bretschneider and Peters (2017); Sanguinetti et al. (2018); Ousidhoum et al. (2019); Mulki et al. (2019); Kurrek et al. (2020); Moon et al. (2020); Mulki and Ghanem (2021); Röttger et al. (2021); Mathew et al. (2021); Shekhar et al. (2022); Toraman et al. (2022); Kirk et al. (2022); Korre et al. (2023); Castillo-lópez et al. (2023); Seo et al. (2024); Sreelakshmi et al. (2024)	20
not reported	Wijesiriwardene et al. (2020)	1

Table 2: Breakdown of datasets by goal.

Languages	Datasets	Count
English	Waseem and Hovy (2016); Waseem (2016); Davidson et al. (2017); Gao and Huang (2017); Jha and Mamidi (2017); Golbeck et al. (2017); Wulczyn et al. (2017); de Gibert et al. (2018); Fersini et al. (2018); ElSherief et al. (2018); Founta et al. (2018); Rezvan et al. (2018); Salminen et al. (2018); Ousidhoum et al. (2019); Zampieri et al. (2019); Borkan et al. (2019); Chung et al. (2019); Qian et al. (2019); Basile et al. (2019); Mandl et al. (2019); Kennedy et al. (2020); Caselli et al. (2020); Pamungkas et al. (2020); Suryawanshi et al. (2020); Wijesiriwardene et al. (2020); Kurrek et al. (2020); Gomez et al. (2020); Vidgen et al. (2020); Pavlopoulos et al. (2020); Raman et al. (2020); Cercas Curry et al. (2021); Vidgen et al. (2021b); Samory et al. (2021); Grimminger and Klinger (2021); Röttger et al. (2021); Pavlopoulos et al. (2021); Fanton et al. (2021); Mathew et al. (2021); Vidgen et al. (2021a); Mollas et al. (2022); Toraman et al. (2022); Kirk et al. (2022); Kennedy et al. (2022); Albanyan et al. (2023); Saker et al. (2023); Sarker et al. (2023); Korre et al. (2023); Park et al. (2023); Kirk et al. (2023); Das et al. (2023); Lee et al. (2024); Yuan and Rizoiu (2025)	55
Italian	Sanguinetti et al. (2018); Bosco et al. (2018); Sprugnoli et al. (2018); Chung et al. (2019); Fersini et al. (2020); Cignarella et al. (2024)	8
German	Roß et al. (2016); Bretschneider and Peters (2017); Wiegand et al. (2019); Mandl et al. (2019); Assenmacher et al. (2021); Demus et al. (2022); Steffen et al. (2023); Goldzycher et al. (2024)	8
Arabic	Mubarak et al. (2017); Albadi et al. (2018); Alakrot et al. (2018); Ousidhoum et al. (2019)	5
Barzilian Portuguese	de Pelle and Moreira (2017); Leite et al. (2020); Trajano et al. (2024)	5
Croatian	Ljubešić et al. (2018); Shekhar et al. (2020, 2022)	4
Spanish, French, Indonesian, Korean (3 each)	Alfina et al. (2017); Fersini et al. (2018); Ibrohim and Budi (2018); Ousidhoum et al. (2019); Chung et al. (2019); Basile et al. (2019); Ibrohim and Budi (2019); Moon et al. (2020); Ollagnier et al. (2022); Park et al. (2023); Castillo-lópez et al. (2023); Seo et al. (2024)	3×4
Hindi, Danish, Turkish, Greek, Russian, Mexican Spanish, Portuguese (2 each)	Pavlopoulos et al. (2017); Álvarez-Carmona et al. (2018); Mandl et al. (2019); Fortuna et al. (2019); Sigurbergsson and Derczynski (2020); Pitenis et al. (2020); Bhardwaj et al. (2020); Zueva et al. (2020); Çöltekin (2020); Zeinert et al. (2021); Saitov and Derczynski (2021); Toraman et al. (2022); Vásquez et al. (2023); Ferreira et al. (2024)	2×7
Slovenian, Levantine, Bengali, Dutch, Albanian, Chinese, Hinglish, Polish, Roman Urdu, Hausa, Ukrainian, Urdu (1 each)	Ljubešić et al. (2018); Mathur et al. (2018); Mulki et al. (2019); Ptaszynski et al. (2019); Rizwan et al. (2020); Romim et al. (2021); Caselli et al. (2021); Nurce et al. (2022); Jiang et al. (2022); Saeed et al. (2023); Vargas et al. (2024); Dementieva et al. (2024)	1×12
Mixed languages	Estonian, Russian: Shekhar et al. (2020); Arabic, Levantine: Mulki and Ghanem (2021); Singlish, Malay, and Tamil: Ng et al. (2024)	3
Code-switched languages	Hindi, English ($n = 6$): Kumar et al. (2018); Bohra et al. (2018); Rawat et al. (2023); Madhu et al. (2023); Singh et al. (2024); Malayalam, English ($n = 1$): Sreelakshmi et al. (2024); Bengali, English ($n = 1$): Raihan et al. (2023); Yoruba, Naija, English ($n = 1$): Ilevbare et al. (2024)	9

Table 3: Breakdown of datasets by language. Datasets labelled as "mixed languages" contain texts from multiple languages, but individual texts are not code-mixed. In contrast, "code-switched datasets" refer to datasets where individual entries exhibit code-switching.

Source	Datasets	Count
Twitter	Waseem and Hovy (2016); Waseem (2016); Mubarak et al. (2017); Davidson et al. (2017); Jha and Mamidi (2017); Golbeck et al. (2017); Roß et al. (2016); Alfina et al. (2017); Albadi et al. (2018); Fersini et al. (2018); ElSherief et al. (2018); Founta et al. (2018); Rezvan et al. (2018); Wiegand et al. (2019); Kumar et al. (2018); Mathur et al. (2018); Bohra et al. (2018); Ibrohim and Budi (2018); Sanguinetti et al. (2018); Bosco et al. (2018); Álvarez-Carmona et al. (2018); Ousidhoum et al. (2019); Mulki et al. (2019); Zampieri et al. (2019); Chung et al. (2019); Basile et al. (2019); Mandl et al. (2019); Ibrohim and Budi (2019); Ptaszynski et al. (2019); Fortuna et al. (2019); Sigurbergsson and Derczynski (2020); Kennedy et al. (2020); Wijesiriwardene et al. (2020); Gomez et al. (2020); Vidgen et al. (2020); Pitenis et al. (2020); Bhardwaj et al. (2020); Fersini et al. (2020); Leite et al. (2020); Çöltekin (2020); Rizwan et al. (2020); Mulki and Ghanem (2021); Zeinert et al. (2021); Caselli et al. (2021); Samory et al. (2021); Grimminger and Klinger (2021); Mathew et al. (2021); Toraman et al. (2022); Kirk et al. (2022); Demus et al. (2022); Albanyan et al. (2023); Trajano et al. (2024); Castillo-lópez et al. (2023); Rawat et al. (2023); Vásquez et al. (2023); Saeed et al. (2023); Madhu et al. (2023); Cignarella et al. (2024); Ilevbare et al. (2024); Ferreira et al. (2024); Yuan and Rizoiu (2025)	70
Facebook	Bretschneider and Peters (2017); Salminen et al. (2018); Kumar et al. (2018); Bosco et al. (2018); Mandl et al. (2019); Sigurbergsson and Derczynski (2020); Bhardwaj et al. (2020); Romim et al. (2021); Zeinert et al. (2021); Raihan et al. (2023); Cignarella et al. (2024); Vargas et al. (2024); Singh et al. (2024)	15
YouTube	Alakrot et al. (2018); Salminen et al. (2018); Kennedy et al. (2020); Romim et al. (2021); Mollas et al. (2022); Nurce et al. (2022); Trajano et al. (2024); Park et al. (2023); Lee et al. (2024); Sreelakshmi et al. (2024)	11
Reddit	Qian et al. (2019); Sigurbergsson and Derczynski (2020); Kennedy et al. (2020); Kurrek et al. (2020); Zeinert et al. (2021); Vidgen et al. (2021a); Mollas et al. (2022); Kirk et al. (2023); Singh et al. (2024); Lee et al. (2024)	10
Instagram	Nurce et al. (2022); Singh et al. (2024)	2
Gab & Stormfront	de Gibert et al. (2018); Qian et al. (2019); Mathew et al. (2021); Kennedy et al. (2022); Kirk et al. (2023)	5
Human Creation	Chung et al. (2019); Cercas Curry et al. (2021); Fanton et al. (2021); Ollagnier et al. (2022); Goldzycher et al. (2024)	7
Synthetic	Vidgen et al. (2021b); Röttger et al. (2021); Kirk et al. (2022)	3
Existing datasets	Caselli et al. (2020); Pamungkas et al. (2020); Pavlopoulos et al. (2021); Saker et al. (2023); Trajano et al. (2024); Korre et al. (2023); Seo et al. (2024); Ng et al. (2024); Dementieva et al. (2024); Lee et al. (2024)	10
Other	Mubarak et al. (2017); Gao and Huang (2017); Wulczyn et al. (2017); Pavlopoulos et al. (2017); de Pelle and Moreira (2017); Ljubešić et al. (2018); Sprugnoli et al. (2018); Borkan et al. (2019); Shekhar et al. (2020); Suryawanshi et al. (2020); Pavlopoulos et al. (2020); Moon et al. (2020); Zueva et al. (2020); Raman et al. (2020); Assenmacher et al. (2021); Saitov and Derczynski (2021); Jiang et al. (2022); Shekhar et al. (2022); Sarker et al. (2023); Steffen et al. (2023); Das et al. (2023)	24

Table 4: Breakdown of datasets by data source.

Collection method	Datasets	Count
Keyword-based	Waseem and Hovy (2016); Mubarak et al. (2017); Davidson et al. (2017); Jha and Mamidi (2017); Golbeck et al. (2017); Roß et al. (2016); Alfina et al. (2017); Albadi et al. (2018); Fersini et al. (2018); ElSherief et al. (2018); Rezvan et al. (2018); Salminen et al. (2018); Wiegand et al. (2019); Kumar et al. (2018); Mathur et al. (2018); Bohra et al. (2018); Ibrohim and Budi (2018); Sanguinetti et al. (2018); Bosco et al. (2018); Álvarez-Carmona et al. (2018); Ousidhoum et al. (2019); Mulki et al. (2019); Zampieri et al. (2019); Qian et al. (2019); Basile et al. (2019); Mandl et al. (2019); Ibrohim and Budi (2019); Fortuna et al. (2019); Sigurbergsson and Derczynski (2020); Pamungkas et al. (2020); Wijesiriwardene et al. (2020); Kurrek et al. (2020); Gomez et al. (2020); Vidgen et al. (2020); Pitenis et al. (2020); Bhardwaj et al. (2020); Leite et al. (2020); Rizwan et al. (2020); Romim et al. (2021); Zeinert et al. (2021); Caselli et al. (2021); Cercas Curry et al. (2021); Samory et al. (2021); Grimminger and Klinger (2021); Mathew et al. (2021); Jiang et al. (2022); Toraman et al. (2022); Kirk et al. (2022); Demus et al. (2022); Trajano et al. (2024); Castillo-lópez et al. (2023); Rawat et al. (2023); Kirk et al. (2023); Vásquez et al. (2023); Raihan et al. (2023); Saeed et al. (2023); Das et al. (2023); Cignarella et al. (2024); Seo et al. (2024); Vargas et al. (2024); Singh et al. (2024); Ferreira et al. (2024); Lee et al. (2024)	73
Keypage-based	Gao and Huang (2017); Bretschneider and Peters (2017); Alakrot et al. (2018); Fersini et al. (2018); Kumar et al. (2018); Bosco et al. (2018); Qian et al. (2019); Sigurbergsson and Derczynski (2020); Kennedy et al. (2020); Kurrek et al. (2020); Raman et al. (2020); Mulki and Ghanem (2021); Romim et al. (2021); Vidgen et al. (2021a); Nurce et al. (2022); Trajano et al. (2024); Park et al. (2023); Kirk et al. (2023); Steffen et al. (2023); Raihan et al. (2023); Cignarella et al. (2024); Ilevbare et al. (2024); Vargas et al. (2024); Singh et al. (2024)	26
Keyuser-based	Waseem and Hovy (2016); Wulczyn et al. (2017); Fersini et al. (2018); ElSherief et al. (2018); Wiegand et al. (2019); Mulki et al. (2019); Basile et al. (2019); Mandl et al. (2019); Ptaszynski et al. (2019); Fortuna et al. (2019); Wijesiriwardene et al. (2020); Kurrek et al. (2020); Leite et al. (2020); Rizwan et al. (2020); Zeinert et al. (2021); Caselli et al. (2021); Nurce et al. (2022); Trajano et al. (2024); Rawat et al. (2023); Singh et al. (2024); Yuan and Rizoiu (2025)	25
Heuristics	ElSherief et al. (2018); Salminen et al. (2018); Kennedy et al. (2020); Albanyan et al. (2023); Sarker et al. (2023); Kirk et al. (2023)	7
Using all available data	Pavlopoulos et al. (2017); Ljubešić et al. (2018); Shekhar et al. (2020); Assenmacher et al. (2021)	7
Geolocation	Mathur et al. (2018); Álvarez-Carmona et al. (2018); Caselli et al. (2021); Castillo-lópez et al. (2023); Vásquez et al. (2023)	5
Other	Mubarak et al. (2017); Wulczyn et al. (2017); de Pelle and Moreira (2017); de Gibert et al. (2018); Founta et al. (2018); Sprugnoli et al. (2018); Moon et al. (2020); Çöltekin (2020); Mollas et al. (2022); Kennedy et al. (2022); Madhu et al. (2023); Ng et al. (2024)	12
not reported	Zueva et al. (2020); Shekhar et al. (2022); Trajano et al. (2024); Sreelakshmi et al. (2024)	4

Table 5: Breakdown of datasets by collection methods.

Task formulation	Datasets	Count
Binary classification	Gao and Huang (2017); Golbeck et al. (2017); Wulczyn et al. (2017); Roß et al. (2016); Pavlopoulos et al. (2017); Alfina et al. (2017); Alakrot et al. (2018); Ljubešić et al. (2018); ElSherief et al. (2018); Bohra et al. (2018); Álvarez-Carmona et al. (2018); Qian et al. (2019); Suryawanshi et al. (2020); Pavlopoulos et al. (2020); Raman et al. (2020); Romim et al. (2021); Assenmacher et al. (2021); Saitov and Derczynski (2021); Kirk et al. (2022); Sarker et al. (2023); Korre et al. (2023); Park et al. (2023); Das et al. (2023); Madhu et al. (2023); Goldzycher et al. (2024); Cignarella et al. (2024); Ilevbare et al. (2024); Dementieva et al. (2024); Ferreira et al. (2024); Lee et al. (2024); Sreelakshmi et al. (2024)	34
Multi-class classification	Waseem and Hovy (2016); Waseem (2016); Mubarak et al. (2017); Davidson et al. (2017); Jha and Mamidi (2017); de Gibert et al. (2018); Rezvan et al. (2018); Mathur et al. (2018); Ibrohim and Budi (2018); Mulki et al. (2019); Caselli et al. (2020); Wijesiriwardene et al. (2020); Kurrek et al. (2020); Gomez et al. (2020); Pitenis et al. (2020); Moon et al. (2020); Leite et al. (2020); Grimminger and Klinger (2021); Toraman et al. (2022); Castillo-lópez et al. (2023); Rawat et al. (2023); Yuan and Rizoiu (2025)	24
Multi-label classification	Founta et al. (2018); Ibrohim and Budi (2019); Shekhar et al. (2022); Kennedy et al. (2022)	4
Hierarchical	Bretschneider and Peters (2017); Pavlopoulos et al. (2017); de Pelle and Moreira (2017); Fersini et al. (2018); Salminen et al. (2018); Wiegand et al. (2019); Kumar et al. (2018); Sanguinetti et al. (2018); Albadi et al. (2018); Bosco et al. (2018); Sprugnoli et al. (2018); Zampieri et al. (2019); Chung et al. (2019); Basile et al. (2019); Mandl et al. (2019); Ptaszynski et al. (2019); Fortuna et al. (2019); Shekhar et al. (2020); Sigurbergsson and Derczynski (2020); Bhardwaj et al. (2020); Çöltekin (2020); Rizwan et al. (2020); Zeinert et al. (2021); Caselli et al. (2021); Cercas Curry et al. (2021); Vidgen et al. (2021b); Röttger et al. (2021); Mathew et al. (2021); Vidgen et al. (2021a); Mollas et al. (2022); Nurce et al. (2022); Jiang et al. (2022); Kirk et al. (2022); Albanyan et al. (2023); Trajano et al. (2024); Kirk et al. (2023); Vásquez et al. (2023); Raihan et al. (2023); Saeed et al. (2023); Vargas et al. (2024); Singh et al. (2024); Ng et al. (2024)	54
Parallel	Ousidhoum et al. (2019); Fersini et al. (2020); Mulki and Ghanem (2021); Steffen et al. (2023); Cignarella et al. (2024)	7
Other	Pamungkas et al. (2020); Zueva et al. (2020); Samory et al. (2021); Pavlopoulos et al. (2021); Ollagnier et al. (2022); Saker et al. (2023)	6
not reported	Zueva et al. (2020); Shekhar et al. (2022); Trajano et al. (2024); Sreelakshmi et al. (2024)	4

Table 6: Breakdown of datasets by task types.

Number of annotators	Datasets	Count
Involving single annotator (partially or fully)	Gao and Huang (2017); Ljubešić et al. (2018); Salminen et al. (2018); Wiegand et al. (2019); Ibrohim and Budi (2019); Pamungkas et al. (2020); Suryawanshi et al. (2020); Çöltekin (2020); Caselli et al. (2021); Vidgen et al. (2021b); Grimminger and Klinger (2021); Ollagnier et al. (2022); Goldzycher et al. (2024); Ferreira et al. (2024)	14
Multiple, subset	Roß et al. (2016); Alfina et al. (2017); Ibrohim and Budi (2018); Bosco et al. (2018); Ibrohim and Budi (2019); Ptaszynski et al. (2019); Fortuna et al. (2019); Pamungkas et al. (2020); Suryawanshi et al. (2020); Kurrek et al. (2020); Vidgen et al. (2020); Leite et al. (2020); Romim et al. (2021); Zeinert et al. (2021); Cercas Curry et al. (2021); Vidgen et al. (2021b); Grimminger and Klinger (2021); Röttger et al. (2021); Shekhar et al. (2022); Toraman et al. (2022); Kirk et al. (2022); Demus et al. (2022); Kirk et al. (2023); Steffen et al. (2023); Raihan et al. (2023); Das et al. (2023); Madhu et al. (2023)	29
Multiple, full set	Golbeck et al. (2017); Bretschneider and Peters (2017); Pavlopoulos et al. (2017); de Pelle and Moreira (2017); Alakrot et al. (2018); Ljubešić et al. (2018); Fersini et al. (2018); Rezvan et al. (2018); Mathur et al. (2018); Bohra et al. (2018); Sprugnoli et al. (2018); Álvarez-Carmona et al. (2018); Mulki et al. (2019); Chung et al. (2019); Caselli et al. (2020); Wijesiriwardene et al. (2020); Pitenis et al. (2020); Rizwan et al. (2020); Mulki and Ghanem (2021); Caselli et al. (2021); Pavlopoulos et al. (2021); Fanton et al. (2021); Vidgen et al. (2021a); Saitov and Derczynski (2021); Nurce et al. (2022); Jiang et al. (2022); Kirk et al. (2022); Kennedy et al. (2022); Albanyan et al. (2023); Saker et al. (2023); Sarker et al. (2023); Park et al. (2023); Castillo-lópez et al. (2023); Rawat et al. (2023); Vásquez et al. (2023); Saeed et al. (2023); Cignarella et al. (2024); Ilevbare et al. (2024); Vargas et al. (2024); Singh et al. (2024); Lee et al. (2024); Sreelakshmi et al. (2024)	47
Involving crowdsourcing	Mubarak et al. (2017); Davidson et al. (2017); Wulczyn et al. (2017); Albadi et al. (2018); Fersini et al. (2018); Founta et al. (2018); Ousidhoum et al. (2019); Zampieri et al. (2019); Borkan et al. (2019); Qian et al. (2019); Basile et al. (2019); Kennedy et al. (2020); Gomez et al. (2020); Pavlopoulos et al. (2020); Samory et al. (2021); Mathew et al. (2021); Mollas et al. (2022); Assenmacher et al. (2021); Kumar et al. (2018); Moon et al. (2020); Korre et al. (2023); Yuan and Rizoiu (2025)	29
not reported or unclear	Waseem and Hovy (2016); Jha and Mamidi (2017); Ljubešić et al. (2018); de Gibert et al. (2018); ElSherief et al. (2018); Bosco et al. (2018); Mandl et al. (2019); Shekhar et al. (2020); Sigurbergsson and Derczynski (2020); Bhardwaj et al. (2020); Fersini et al. (2020); Zueva et al. (2020); Raman et al. (2020); Trajano et al. (2024); Seo et al. (2024); Ng et al. (2024); Dementieva et al. (2024)	23

Table 7: Breakdown of datasets by numbers of annotators.

Reported Demographics	**Datasets**	**Count**
Age	Roß et al. (2016); Alfina et al. (2017); Alakrot et al. (2018); Founta et al. (2018); Chung et al. (2019); Ibrohim and Budi (2019); Sigurbergsson and Derczynski (2020); Kurrek et al. (2020); Vidgen et al. (2020); Leite et al. (2020); Zeinert et al. (2021); Caselli et al. (2021); Cercas Curry et al. (2021); Vidgen et al. (2021b); Grimminger and Klinger (2021); Röttger et al. (2021); Vidgen et al. (2021a); Assenmacher et al. (2021); Saitov and Derczynski (2021); Nurce et al. (2022); Toraman et al. (2022); Kirk et al. (2022, 2023); Vásquez et al. (2023); Raihan et al. (2023); Goldzycher et al. (2024); Ng et al. (2024); Lee et al. (2024)	33
Gender	Roß et al. (2016); Alfina et al. (2017); Founta et al. (2018); Chung et al. (2019); Ibrohim and Budi (2019); Sigurbergsson and Derczynski (2020); Suryawanshi et al. (2020); Kurrek et al. (2020); Vidgen et al. (2020); Leite et al. (2020); Mulki and Ghanem (2021); Zeinert et al. (2021); Caselli et al. (2021); Cercas Curry et al. (2021); Vidgen et al. (2021b); Grimminger and Klinger (2021); Röttger et al. (2021); Vidgen et al. (2021a); Assenmacher et al. (2021); Saitov and Derczynski (2021); Nurce et al. (2022); Jiang et al. (2022); Kirk et al. (2022); Saker et al. (2023); Vásquez et al. (2023); Raihan et al. (2023); Goldzycher et al. (2024); Ilevbare et al. (2024); Ng et al. (2024); Lee et al. (2024)	33
Language	Gao and Huang (2017); Albadi et al. (2018); Rezvan et al. (2018); Wiegand et al. (2019); Ousidhoum et al. (2019); Chung et al. (2019); Sigurbergsson and Derczynski (2020); Vidgen et al. (2020); Çöltekin (2020); Mulki and Ghanem (2021); Zeinert et al. (2021); Caselli et al. (2021); Cercas Curry et al. (2021); Vidgen et al. (2021b); Grimminger and Klinger (2021); Röttger et al. (2021); Saitov and Derczynski (2021); Nurce et al. (2022); Kirk et al. (2022); Castillo-lópez et al. (2023); Kirk et al. (2023); Vásquez et al. (2023); Raihan et al. (2023); Goldzycher et al. (2024); Vargas et al. (2024); Ng et al. (2024)	32
Education	Founta et al. (2018); Chung et al. (2019); Ibrohim and Budi (2019); Vidgen et al. (2020); Romim et al. (2021); Caselli et al. (2021); Cercas Curry et al. (2021); Vidgen et al. (2021b); Grimminger and Klinger (2021); Röttger et al. (2021); Toraman et al. (2022); Kirk et al. (2022); Rawat et al. (2023); Kirk et al. (2023); Vásquez et al. (2023); Raihan et al. (2023); Das et al. (2023); Goldzycher et al. (2024); Ilevbare et al. (2024); Vargas et al. (2024); Ng et al. (2024); Lee et al. (2024)	27
Location (nationality, country of origin, IP)	Mubarak et al. (2017); Albadi et al. (2018); Alakrot et al. (2018); Founta et al. (2018); Mulki et al. (2019); Vidgen et al. (2020, 2021b); Röttger et al. (2021); Vidgen et al. (2021a); Assenmacher et al. (2021); Nurce et al. (2022); Kirk et al. (2022); Castillo-lópez et al. (2023); Kirk et al. (2023); Vásquez et al. (2023); Vargas et al. (2024)	18
Race and ethnicity	Alfina et al. (2017); Ibrohim and Budi (2019); Sigurbergsson and Derczynski (2020); Kurrek et al. (2020); Leite et al. (2020); Zeinert et al. (2021); Caselli et al. (2021); Cercas Curry et al. (2021); Vidgen et al. (2021b); Röttger et al. (2021); Vidgen et al. (2021a); Kirk et al. (2022, 2023); Lee et al. (2024)	16
not reported	Waseem and Hovy (2016); Waseem (2016); Mubarak et al. (2017); Davidson et al. (2017); Jha and Mamidi (2017); Golbeck et al. (2017); Wulczyn et al. (2017); Bretschneider and Peters (2017); Pavlopoulos et al. (2017); de Pelle and Moreira (2017); Ljubešić et al. (2018); de Gibert et al. (2018); Fersini et al. (2018); ElSherief et al. (2018); Salminen et al. (2018); Kumar et al. (2018); Mathur et al. (2018); Bohra et al. (2018); Ibrohim and Budi (2018); Sanguinetti et al. (2018); Bosco et al. (2018); Sprugnoli et al. (2018); Álvarez-Carmona et al. (2018); Zampieri et al. (2019); Borkan et al. (2019); Qian et al. (2019); Basile et al. (2019); Mandl et al. (2019); Ptaszynski et al. (2019); Fortuna et al. (2019); Shekhar et al. (2020); Kennedy et al. (2020); Caselli et al. (2020); Pamungkas et al. (2020); Wijesiriwardene et al. (2020); Gomez et al. (2020); Pavlopoulos et al. (2020); Pitenis et al. (2020); Bhardwaj et al. (2020); Moon et al. (2020); Fersini et al. (2020); Zueva et al. (2020); Rizwan et al. (2020); Raman et al. (2020); Samory et al. (2021); Pavlopoulos et al. (2021); Fanton et al. (2021); Mathew et al. (2021); Mollas et al. (2022); Shekhar et al. (2022); Ollagnier et al. (2022); Demus et al. (2022); Albanyan et al. (2023); Sarker et al. (2023); Trajano et al. (2024); Korre et al. (2023); Park et al. (2023); Madhu et al. (2023); Cignarella et al. (2024); Seo et al. (2024); Dementieva et al. (2024); Singh et al. (2024); Ferreira et al. (2024); Yuan and Rizoiu (2025)	78

Table 8: Examples of annotator demographics and datasets that report them.

Methods to resolve disagreements	Datasets	Count
Majority vote	Samory et al. (2021); Qian et al. (2019); Fortuna et al. (2019); Rezvan et al. (2018); Wijesiriwardene et al. (2020); Waseem (2016); Davidson et al. (2017); Moon et al. (2020); Shekhar et al. (2022); Alakrot et al. (2018); Pavlopoulos et al. (2017); Sreelakshmi et al. (2024); Saeed et al. (2023); Demus et al. (2022); Mathur et al. (2018); Wulczyn et al. (2017); Lee et al. (2024); Gomez et al. (2020); Korre et al. (2023); Romim et al. (2021); Mathew et al. (2021); Vargas et al. (2024); Vásquez et al. (2023); Caselli et al. (2020); Kennedy et al. (2022); Mulki et al. (2019); Founta et al. (2018); Toraman et al. (2022); Mulki and Ghanem (2021); Ibrohim and Budi (2019); Ousidhoum et al. (2019); Suryawanshi et al. (2020); de Pelle and Moreira (2017); Pitenis et al. (2020); Trajano et al. (2024); Fersini et al. (2018); ElSherief et al. (2018); Zampieri et al. (2019); Basile et al. (2019); Pavlopoulos et al. (2021)	48
Additional annotators	Golbeck et al. (2017); Fersini et al. (2018); Mathur et al. (2018); Sanguinetti et al. (2018); Zampieri et al. (2019); Basile et al. (2019); Ptaszynski et al. (2019); Pamungkas et al. (2020); Kurrek et al. (2020); Vidgen et al. (2020); Çöltekin (2020); Vidgen et al. (2021a); Toraman et al. (2022); Kirk et al. (2022); Castillo-lópez et al. (2023); Rawat et al. (2023); Kirk et al. (2023); Raihan et al. (2023); Das et al. (2023); Madhu et al. (2023); Goldzycher et al. (2024); Cignarella et al. (2024)	27
Moderation meeting	Salminen et al. (2018); Zeinert et al. (2021); Caselli et al. (2021); Albanyan et al. (2023); Saker et al. (2023); Sarker et al. (2023); Ilevbare et al. (2024); Ferreira et al. (2024)	8
Other	Gao and Huang (2017); Bretschneider and Peters (2017); Albadi et al. (2018); Alakrot et al. (2018); Pavlopoulos et al. (2020); Leite et al. (2020); Assenmacher et al. (2021); Trajano et al. (2024); Vásquez et al. (2023); Yuan and Rizoiu (2025)	12
Discarded	Davidson et al. (2017); Alfina et al. (2017); Albadi et al. (2018); Ibrohim and Budi (2018); Mulki et al. (2019); Ibrohim and Budi (2019); Rizwan et al. (2020); Mulki and Ghanem (2021); Mathew et al. (2021)	9
not applicable	Roß et al. (2016); Ljubešić et al. (2018); Wiegand et al. (2019); Chung et al. (2019); Shekhar et al. (2020); Kennedy et al. (2020); Vidgen et al. (2021b); Grimminger and Klinger (2021); Röttger et al. (2021); Ollagnier et al. (2022); Ng et al. (2024)	16
not reported	Waseem and Hovy (2016); Mubarak et al. (2017); Jha and Mamidi (2017); de Gibert et al. (2018); Kumar et al. (2018); Bohra et al. (2018); Bosco et al. (2018); Sprugnoli et al. (2018); Álvarez-Carmona et al. (2018); Borkan et al. (2019); Mandl et al. (2019); Sigurbergsson and Derczynski (2020); Bhardwaj et al. (2020); Fersini et al. (2020); Zueva et al. (2020); Raman et al. (2020); Fanton et al. (2021); Mollas et al. (2022); Saitov and Derczynski (2021); Nurce et al. (2022); Jiang et al. (2022); Park et al. (2023); Seo et al. (2024); Dementieva et al. (2024); Singh et al. (2024)	31

Table 9: Breakdown of datasets by label aggregation strategies.

Methods to resolve disagreements	Datasets	Count
Metrics-based selection (crowdsourcing)	ElSherief et al. (2018); Ousidhoum et al. (2019); Qian et al. (2019); Samory et al. (2021); Mathew et al. (2021); Assenmacher et al. (2021); Yuan and Rizoiu (2025)	10
Training	Golbeck et al. (2017); Kurrek et al. (2020); Vidgen et al. (2020, 2021b,a); Shekhar et al. (2022); Kennedy et al. (2022); Demus et al. (2022); Trajano et al. (2024); Vásquez et al. (2023); Dementieva et al. (2024)	13
Moderation meetings only to resolve disagreements	Gao and Huang (2017); Golbeck et al. (2017); Caselli et al. (2020); Kurrek et al. (2020); Zeinert et al. (2021); Caselli et al. (2021); Cercas Curry et al. (2021); Kirk et al. (2022); Ollagnier et al. (2022); Demus et al. (2022); Vásquez et al. (2023); Das et al. (2023)	12
Moderation meetings to refine guidelines	Kumar et al. (2018); Suryawanshi et al. (2020); Jiang et al. (2022); Kirk et al. (2023); Park et al. (2023); Raihan et al. (2023); Cignarella et al. (2024)	10
Tests (During onboarding or hidden during annotation)	Wulczyn et al. (2017); ElSherief et al. (2018); Albadi et al. (2018); Zampieri et al. (2019); Basile et al. (2019); Samory et al. (2021); Mollas et al. (2022); Assenmacher et al. (2021); Korre et al. (2023); Kirk et al. (2023); Lee et al. (2024)	12
Validation by outside annotators	Waseem and Hovy (2016); Jha and Mamidi (2017); Salminen et al. (2018); Romim et al. (2021); Vidgen et al. (2021b); Röttger et al. (2021); Goldzycher et al. (2024); Ptaszynski et al. (2019); Ilevbare et al. (2024); Dementieva et al. (2024)	10
not reported or unclear	Waseem (2016); Mubarak et al. (2017); Davidson et al. (2017); Roß et al. (2016); Bretschneider and Peters (2017); Alfina et al. (2017); de Pelle and Moreira (2017); Alakrot et al. (2018); Ljubešić et al. (2018); Founta et al. (2018); Rezvan et al. (2018); Mathur et al. (2018); Bohra et al. (2018); Ibrohim and Budi (2018); Bosco et al. (2018); Sprugnoli et al. (2018); Álvarez-Carmona et al. (2018); Mulki et al. (2019); Borkan et al. (2019); Chung et al. (2019); Basile et al. (2019); Mandl et al. (2019); Ibrohim and Budi (2019); Fortuna et al. (2019); Shekhar et al. (2020); Sigurbergsson and Derczynski (2020); Pamungkas et al. (2020); Wijesiriwardene et al. (2020); Pavlopoulos et al. (2020); Pitenis et al. (2020); Bhardwaj et al. (2020); Moon et al. (2020); Fersini et al. (2020); Leite et al. (2020); Zueva et al. (2020); Çöltekin (2020); Rizwan et al. (2020); Raman et al. (2020); Mulki and Ghanem (2021); Pavlopoulos et al. (2021); Fanton et al. (2021); Saitov and Derczynski (2021); Nurce et al. (2022); Toraman et al. (2022); Kirk et al. (2022); Albanyan et al. (2023); Saker et al. (2023); Sarker et al. (2023); Castillo-lópez et al. (2023); Rawat et al. (2023); Steffen et al. (2023); Saeed et al. (2023); Madhu et al. (2023); Cignarella et al. (2024); Seo et al. (2024); Vargas et al. (2024); Ng et al. (2024); Singh et al. (2024); Ferreira et al. (2024); Sreelakshmi et al. (2024)	70

Table 10: Breakdown of datasets by quality assurance steps.