From Granular Grief to Binary Belief: A Collaborative Optimization of Annotation Techniques for Anti-Autistic Language

NABA RIZVI, University of California, San Diego, USA
ALEXIS MORALES-FLORES, University of California, San Diego, USA
MOHAMMAD RIZVI, Georgia Institute of Technology, USA
NEDJMA OUSIDHOUM, Cardiff University, Wales
IMANI MUNYAKA, University of California, San Diego, USA

Annotating text for subjective tasks, such as labeling ableist and anti-autistic texts, is a challenge that has attracted significant attention as commonly adopted annotation paradigms, e.g., using majority voting, fall short in capturing the nuances of hate speech or bias annotations. Labeling ableist and anti-autistic texts presents similar challenges in addition to the need for familiarity with autism and anti-autistic discrimination. In this paper, we adopt a collaborative and annotator-centric approach to study the impact of various annotation techniques. We recruit 6 participants to annotate sets of sentences from our 11,596 sentence corpus. The groups annotate through schemes focused on score-based classification, algorithmic labeling, and comparison-based labeling to identify instances of anti-autistic ableist speech. As a result of changes in annotation schemes, our annotator groups shift from a worse-than-chance agreement to moderate agreement. This suggests that implementing annotator group discussion and collecting annotator feedback is likely to result in improved agreement scores in difficult and highly subjective tasks. Our results highlight the importance of a collaborative approach in highly subjective classification tasks as it may lead to an improved understanding of their own biases, and large improvements in agreement scores, particularly among annotators with higher rates of disagreement. **Warning:** *This paper contains examples that may be offensive or upsetting, including explicit slurs used against people with disabilities.*

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Computing methodologies** → **Language resources**.

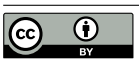
Additional Key Words and Phrases: Autism, Hate Speech, Ableism, Annotators, Participatory Design

ACM Reference Format:
Naba Rizvi, Alexis Morales-Flores, Mohammad Rizvi, Nedjma Ousidhoum, and Imani Munyaka. 2025. From Granular Grief to Binary Belief: A Collaborative Optimization of Annotation Techniques for Anti-Autistic Language. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW297 (November 2025), 23 pages. https://doi.org/10.1145/3757478

1 Introduction

As natural language AI technologies become more ubiquitous in our society, it is important to ensure that they do not marginalize historically ignored populations, including the global population of autistic people, which currently exceeds 75 million [50]. Mitigating such biases in AI has been

Authors' Contact Information: Naba Rizvi, University of California, San Diego, La Jolla, CA, USA, nrizvi@ucsd.edu; Alexis Morales-Flores, University of California, San Diego, La Jolla, CA, USA; Mohammad Rizvi, Georgia Institute of Technology, Atlanta, GA, USA; Nedjma Ousidhoum, Cardiff University, Cardiff, Wales; Imani Munyaka, University of California, San Diego, La Jolla, CA, USA.



ACM 2573-0142/2025/11-ARTCSCW297
https://doi.org/10.1145/3757478

explored by prior work at CSCW. This includes understanding and mitigating cognitive biases in AI [7], and investigating methods to enhance participatory AI design by better integrating the perspectives of the impacted stakeholders [87].

AI biases may harm marginalized communities in various ways. For instance, while AI tools such as chatbots are being increasingly used in the recruitment process [42], and are known to make negative assumptions concerning disabilities [24], which are reflected in their biases against candidates with disabilities [49]. Therefore, it is imperative to identify and mitigate existing inequities perpetuated by AI which include using language affiliated with disability stigma [11]. The dynamic nature of such language and the specialized knowledge required of anti-autistic stigma and discrimination, make detecting anti-autistic ableism a subjective and particularly challenging task [4, 85]. As this task has many practical applications such as content moderation which are becoming increasingly automated [43], it is important to ensure such systems do not contribute to the broader marginalization of autistic people in our society [25, 46, 47]. Prior work has demonstrated several limitations of such classification systems, such as reflecting social biases perpetuated in the annotation process [18] and performing poorly in regards to fairness and bias [47]. Additionally, existing models are less sensitive to recognizing hateful speech directed at autistic people, and tend to over-classify disability-related text as 'toxic', which may lead to unfair censorship of community perspectives [48].

Since building high-quality datasets is paramount to the construction of more efficient anti-autistic detection models, we focus on how this is achieved by improving the data collection and annotation processes. Even though a systematic assessment of dataset annotation quality management found that using collaborative approaches contributes to the improvement of overall dataset quality [41] , there is a notable gap in literature exploring such methods, as previous work at CSCW has focused mainly on the outputs of AI models, rather than the data itself [7, 33, 70, 83, 87], and explored other ethical complexities surrounding the data annotation process. These works have looked at the hidden labors that go into developing data intensive AI systems [14], and other tensions that arise when human judgement is used to train and fine-tune models [16]. For example, as these processes may oversimplify the contexts being annotated, they have limitations when used in the real-world [16].

The annotation process is complex by design and can become harder when dealing with subjective tasks such as hate speech and bias annotations. For example, prior work has found there is no universally accepted way to talk about autism [40]. Although 87% of autistic adults in the US prefer identity-first language, these preferences are dynamic and subject to change based on the time period, cultural context, and other factors and often reflect changes to the connotations of words and phrases [74]. This dynamic and diverse preference in language can make annotation tasks and agreement challenging. Traditionally researchers have relied on majority voting in these annotation tasks [2, 41], which may overlook multiple important perspectives and weigh expert and less informed annotations equally. Recently, there has been a rising interest in the creation of alternative paradigms, and models that better reflect this complexity by accepting disagreements as a feature [12].

However, not all disagreements are equal as some are inevitable and some should be avoided. That is, some disagreements may signal a lack of clarity in the guidelines and can be leveraged to improve task modeling, or may be the result of linguistically debatable cases [41, 57]. This leads us to answer the following research questions:

- What challenges do annotators face that lead to disagreements when labeling anti-autistic ableist speech?

- How will various annotation processes impact the perspectives of the annotators and dataset creators toward tasks such as anti-autistic ableist speech detection?

In this paper, we thoroughly examine different annotation strategies of anti-autistic ableist language, which is subject to a high disagreement. Therefore, through an iterative and collaborative annotator-centric design process, we refine our labeling schemes and annotation strategies and examine the thought processes of our annotators and the specific challenges they face that impact their agreement. We engage six annotators in four annotation iterations. Each iteration is followed by a group re-labeling task to generate discussions around disagreements. In the first three iterations, we make adjustments to the granularity of the labels based on annotator feedback, while the fourth round tests techniques that label sentences based on 1) the labels designed by our annotators in the third iteration, 2) sentence comparisons, and 3) a black-box scoring algorithm that assigns labels based on our annotators' responses to a set of questions. Finally, we conduct a virtual co-design session with pre-selected teams where annotators create their own annotation techniques.

While our work specifically focuses on anti-autistic ableism, our findings may be beneficial for other researchers interested in exploring the annotation processes of other similarly difficult subjective tasks such as hate speech, biased language, and affect in general.

2 Related Work

Research focusing on hate speech detection often includes providing benchmark datasets that can be used to train models, develop models for hate speech detection, and analyze hate speech datasets for fairness or bias. This work often focuses on hate speech towards genders [45] and racial groups [30]. However, hate speech towards disabled persons is relatively under-explored [51, 79], leading to issues with classifiers performing poorly when recognizing ableist language [47], and high levels of bias against people with disabilities in toxicity and sentiment analysis models [48].

This social bias, which can further marginalize a community, can influence how people view them and, thus, how annotators label data and models respond to training. Prior work has found that large pre-trained language models generate content containing harmful biases against marginalized communities [52].

2.1 Bias in AI Systems

Prior work suggests that word embeddings perpetuate gender bias present in the text corpus used leading to associating women with being a nurse or other roles, even after debiasing [27]. Buolamwini and Gebru demonstrate that due to the overrepresentation of white male faces in training data, facial recognition technologies underperform on darker-skinned individuals [10]. Such biases extends to hate speech detection, as research indicates sentences written in African-American English are more likely to be incorrectly labeled as offensive by detection models trained on various datasets due to the lack of diversity in the content collected, and labeling regardless of context [19]. Similarly, researchers have shown that toxicity models often label sentences negatively for the mere presence of gender identity words [53] or disability terms [48]. For example, "I am a person with mental illnesses" was rated as 7 times more toxic than "I am a person", thus highlighting how non-inclusive training on people with disabilities can have harmful effects on the entire annotation process and resulting detection models [34].

2.2 Annotation Process

The annotation process can reinforce biases as different demographic groups label posts differently [22, 75]. However, inclusive training can improve understanding of disability and ableism

and potentially have the same impact for annotators [34, 62]. Research shows that the way people judge ableist actions is determined by the gender of the victim and their disability [78]. In particular, judgments about how people with autism are treated are sometimes justified using social bias, where autism is defined solely through the idea that it is a 'deficit' of certain skills that technology can detect or remedy [9, 63, 77]. This has affected the way research is done for that community [62, 63, 71–73, 88] due to the impact of these normative stereotypes [5, 18].

2.3 Disagreement Management

Despite prior issues, hate speech studies continue to use annotators who may not be familiar with the subject area and are influenced by stereotypes or may not consider annotator diversity [35], which can cause low agreement and bias the ground truth based on majority vote [67]. For example, prior work highlighted the differences in perspectives of annotators who had lived experience with the text topic compared to graduate student annotators without this experience [56]. So, in addition to introducing training, researchers are working to identify the source of disagreements and potential solutions to avoid using a majority vote for subjective tasks [64]. Disagreements can emerge from differences in content interpretations. Wan et al. developed a disagreement prediction mechanism that uses annotator demographics and text content to determine where annotators may disagree [81].

Prior work at CSCW has explored ways to handle similar disagreements in collaborative systems in ways that preserve the diverse perspectives of the collaborators. This includes emphasizing that both individual thoughts and group interactions are important as they help team members become aware of each other's perspectives, which can improve coordination [58]. Other work focused on contested collective intelligence by using tools and technology to help people work together and better understand each other's different perspectives [44]. They found that both the machine and human annotations brought unique findings as the machine was better at handling larger amounts of information while the humans were better at making connections [44]. As well, other work has explored methods to effectively manage dissent in online community moderation, as it may sometimes lead to new community standards and values [86].

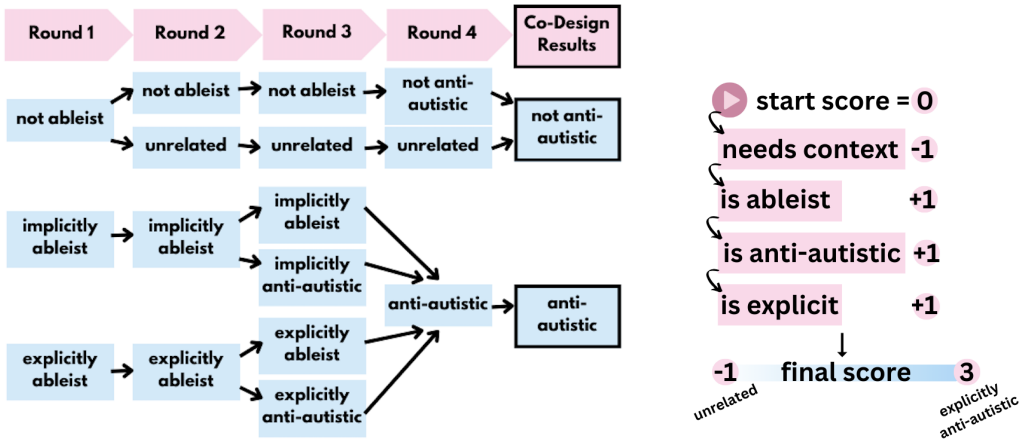
We support this prior work by identifying disagreements through consultation and participatory design, an underutilized technique for identifying sources of disagreement, and annotator mental models for hate speech labeling. Our collaborative approach increases our annotators' understanding of each others' perspectives, thus providing us with insights on their unique thought processes and the differences in their interpretations of labels.

3 Methods

In an effort to determine the barriers annotators may experience while annotating anti-autistic speech, we observe a group annotating an anti-autistic dataset in real time. The annotation process is done in four rounds, followed by a co-design session. We use ethnographic observation to center the annotator's experiences in our results. This section describes the methodologies used to carry out our annotation study. We focus on both the labeling process and the feedback from annotators about the process.

3.1 Participants

This study is led by an autistic researcher. There are six annotators and one of them is a participant-observer who provides an insider perspective during our analysis. One annotator is a computer scientist who is an expert in text-based hate speech detection and the remaining annotators were graduate students at American universities. The number of annotators involved is similar to that of prior studies [1, 29]. All annotators received training on autism before their task assignment



(a) (b)

Fig. 1. Presenting the labeling schemes (a) used in all the rounds for score-based labeling and the logic of our black-box algorithm (b), used in round 4 to dynamically assign scores by asking the annotators a series of questions about each sentence. In our co-design session, we used the same labeling scheme as round 4 which was further simplified by annotators as shown under Co-Design Results in (a).

through the annotation guidelines available here [1]. Four of our annotators are neurodivergent, three were women, and all six were from different racial and ethnic backgrounds. Our training included information on terms and concepts relating to autism and a glossary B of commonly used terms in the online discourse surrounding autism from organizations and activists in addition to scientific literature [3, 28, 55, 68, 89].

3.2 Ethics

Our study is approved by the internal review board at our institution. Participants are warned of the sensitive nature of the data and have access to the services at our institution if needed. The data for the study came from publicly available datasets [26, 76]. However, in an effort to protect post authors, we reword each example post shown in the paper so that it cannot be traced back to a single author.

3.3 Data Collection

We select sentences from Sentiment140, a Twitter dataset, and a dataset released from Reddit in 2015 [26, 76] based on the presence of the following keywords: autism, autistic, r*tard. We pre-process the data by removing duplicates, re-shared posts, non-English sentences, and removing posts containing any media such as images or URLs, and we use the resulting corpus (n = 11596) to pull random sets of sentences for the annotators to label. In order to minimize the possibility of artificial inflation in agreement scores, we use different sets of sentences for each round of our annotation process, and ensure we collect data from the participants individually and in different group settings.

Round 1: Separating Implicit and Explicit Speech.

[1] https://figshare.com/s/b7c122c3cda7342b3651

3.4 Annotation Process

We conduct six rounds of annotation to test the performance of various labeling schemes quantifying anti-autistic sentiments through classifications of their explicitness, ableism, and relevance. The labels we use in these rounds are shown in Figure 1. Each round is concluded with a group discussion on sentences with high rates of disagreements. These scores are not shared with the annotators to avoid biased results during the discussions. The new labels introduced in each round (shown in Figure 1) are based on the feedback received from the annotators during our discussions in the prior round, which we detail below.

Round 1: Separating Implicit and Explicit Speech. We begin by assessing the explicitness of anti-autistic speech. We define "explicit" ableist speech as sentences that contain slurs, violent language, or harmful stereotypes. The "implicitly ableist" category includes all other ableist sentences, including those describing autism as a disease [37]. We create a separate category for "not ableist" sentences.

Round 2: Adding the "Unrelated" Category. We include a new category for 'unrelated' sentences that encompasses sentences that may be incomplete, entirely unrelated to autism, or incomprehensible.

Round 3: Separating Ableism and Anti-Autistic Speech. We separate anti-autistic and ableist sentences from each other. The "ableist" category includes sentences that contain ableist language or slurs that are not exclusively directed at autistic people. The anti-autistic category contains only sentences that are targeting autistic people. We maintain the distinction between implicit and explicit speech employed in rounds 1 and 2.

Round 4: Consolidating Categories. We drop the ableist label with all sentences being classified as simply 'anti-autistic' or not. We remove the distinction between implicit and explicit sentences, but preserve the 'unrelated' label.

Round 5: Black-box Annotation. In this approach, the annotators are not made aware of the underlying algorithm used for scoring, as shown in Figure 1. While the prior labeling schemes gave annotators the freedom to think through the classification of each sentence using their own unique strategies, the black-box technique has a series of pre-defined questions to guide the annotators' thought processes. The starting score for each sentence is set at 0. The final score represents the 'intensity' of anti-autistic sentiments expressed in the sentence. A score of -1 indicates that the sentence is unrelated or needs more context, a score of 0 indicates the sentence is not ableist, and a score of 3 indicates that the sentence is explicitly anti-autistic. Sentences that are implicitly ableist, or ableist but not necessarily anti-autistic will have scores in between 0 and 3.

Round 6: Comparison Annotation. This technique involves creating pairs of sentences, which are then displayed to annotators in a random order to avoid response bias. The annotators are tasked with assessing which sentence in each pair is more ableist. This method allows for the examination of anti-autistic ableism on a spectrum, providing a comparative view of the severity of ableist sentiments between sentences.

3.5 Labels

Our annotation instructions include a description of task steps and label definitions with examples. The full document provided to annotators can be found in Appendix A. Annotators are asked to classify sentences as either i) implicitly or ii) explicitly ableist toward autistic people, or iii) not ableist. These definitions are updated accordingly for each annotation task. The full changes are

available in our Appendix. The following definitions are used in rounds 3 and 4 of our annotation process:

Not-Ableist: Sentences unrelated to autism or disabilities or written by an autistic person reaching out for help and support.

Implicitly Ableist: Making assumptions or comments about a disabled person's abilities that would not be made about an able-bodied person, which includes inspiration porn, use of condescending language, and infantilization [31].

Implicitly Anti-Autistic: Sentences that describe autism as an "illness" or "disease", or focus on clinical applications such as "curing" or "diagnosing" autism [9, 61].

Explicitly Ableist: Sentences using slurs for disabled people such as: r*tard, lame, insane, deluded, moron [23].

Explicitly Anti-Autistic: Sentences that dehumanize autistic people, contain ableist slurs such as r*tard, promote negative stereotypes, or express negative emotions such as fear, disgust, or hatred toward autistic people [23].

Unrelated/Needs More Context: text that is completely unrelated to disabilities, needs more context, and/or contains forms of media other than text.

3.6 Co-Design Session

We conduct the co-design session via Zoom after the last round with the participants from round 4 to 6. The session begins with a short discussion on the round 4 sentences, where the annotators individually annotated the same sentences using the labeling scheme detailed in Figure 1. Participants are then asked to individually re-label sentences, with a researcher observing to keep track of time. After this, participants are divided into breakout rooms where they relabel sentences collaboratively with a partner and co-design alternative labels. This is followed by another relabeling session with a different partner. The session concludes with a final discussion involving all participants. We provide the participants with a virtual whiteboard via Google Jamboard containing sentences to relabel on different slides, with the last few slides providing them sentences they can refer to while designing their preferred annotation schemes.

3.7 Data Analysis

We measure the agreement among our annotators using Cohen's Kappa scores [59] and Krippendorff's alpha [32]. The researchers observe the participants during all of the disagreement discussions and take field notes in an unstructured manner. The notes focus on why each label was selected by a particular annotator, and any feedback on specific labels. The researchers may ask the annotators clarifying questions as needed to gain a better understanding of the specific label(s) or sentences that may be points of contention, or the thought process of each annotator to include in their field notes. Through this process of iterative member checking, we seek to have a more nuanced and accurate understanding of the participants' labeling experiences [17].

After collecting field notes from the observations, we conduct an interpretive phenomenological analysis of the data collected [69] by engaging in a structured discussion to analyze the participants' feedback and systematically evaluating each label. One of the researchers provides an insider perspective as they participated in the annotation process. The researchers critically examine whether any labels are missing, redundant, or require refinement. The goal of this collaborative analysis is to develop a detailed list of actionable changes for each label based on the participants' feedback. For example, if any label receives negative feedback from a participant, it is flagged for removal or significant modification in subsequent iterations. Through this approach, we refine

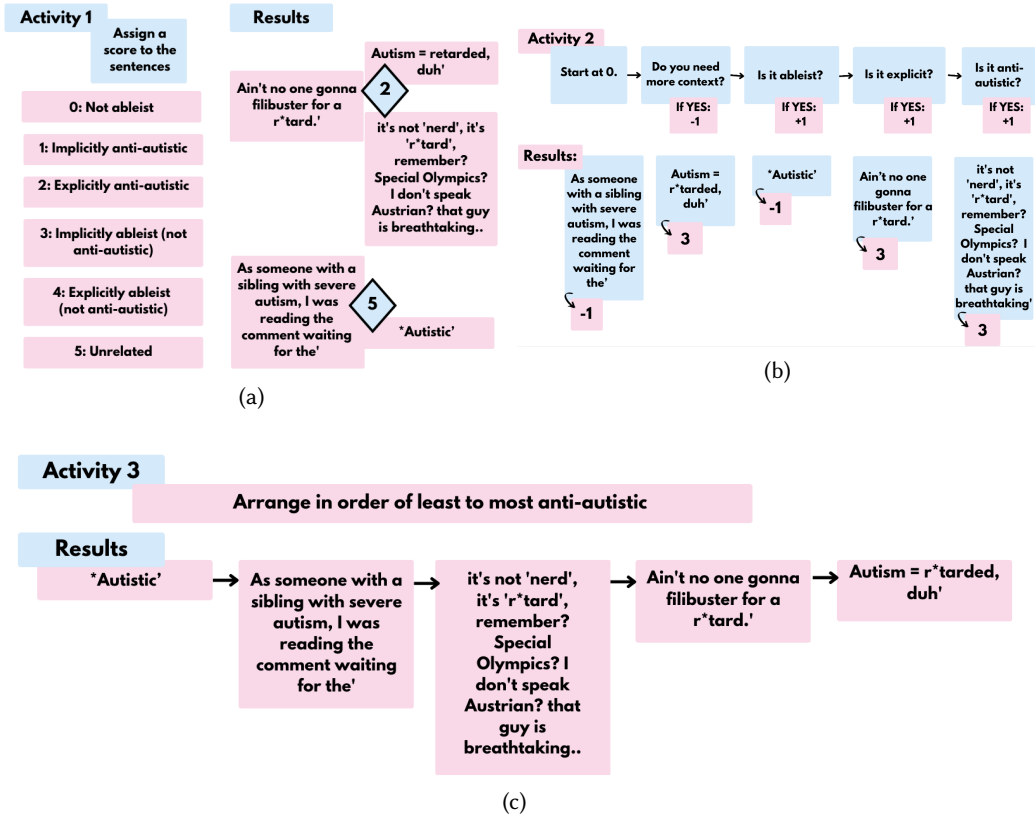


Fig. 2. The collaborative re-labeling activities our annotators complete in our co-design session include assigning labels based on a static scale (a), assigning labels based on our scoring algorithm (b), and arranging sentences in order (c). These activities are designed to replicate the scale-based, black-box, and comparison-based annotation techniques they completed independently in round 4, before our co-design session.

the labeling scheme over successive rounds, ensuring it aligns more closely with the participants' lived experiences and perspectives as identified during the observations. Thus, our labeling scheme evolves dynamically, driven by both the data and ongoing interpretive analysis.

4 Results

In this section, we present our findings from the annotation process, highlighting annotator labeling strategies, sources of disagreement, and outcomes from co-design sessions. Through discussions, we identify factors influencing perceptions of anti-autistic ableism, such as speaker identification, explicit slurs, and key term definitions. We examine disagreements and how iterative updates to labeling schemes can better address the annotators' needs and refine the process.

4.1 Annotation Challenges

Our results reveal that the challenges faced by annotators arise due to the complexity of the labeling schemes, language ambiguity, missing context, and difference in each individual annotators' labeling approach. We also uncover that while labeling collaboratively, the annotators feel less confident in labeling sentences as "not ableist". In order to address these challenges, we iteratively refine the

labeling schemes and definitions introduced in our guidelines to better accommodate annotators' needs.

4.1.1 Granularity of Labeling Schemes. While annotators often search for the presence of explicit slurs, they disagree on whether words such as *r*tard* should be classified as "ableist" or "anti-autistic". Therefore, the definitions of these key terms and their corresponding labeling schemes are iteratively updated– notably increasing in granularity until other annotation techniques beyond the labeling scale are introduced. During the first 3 iterations, annotators believe a more granular scale may help simplify the task by creating clearly defined and highly specific categories as opposed to broader labels, which are more difficult to assess. However, following the co-design session, the annotators reach an agreement that a binary classification of sentences as anti-autistic or not, wherein sentences needing more context are classified as "not anti-autistic" will greatly simplify the annotation task and decrease disagreement.

4.1.2 Specialized Language. The participants highlight semantic concerns arising from the usage of language such as slang, technical terms, or specialized language unique to a particular field (e.g. tweets discussing specific legal codes). For example:

> " Autism HB451 only covers those having insurance that must follow state mandates. How many of us are still left out?"

This sentence requires the annotators to be familiar with the specified house bill of a particular state. Thus, they disagree on whether this sentence should be classified as "not ableist" or "unrelated".

Due to the specialized nature of such language, participants report having to search for terms and concepts before assigning labels. The participants note that some of the sentences were in languages other than English or contain special characters in unicode which are difficult to comprehend.

4.1.3 Missing Context. A lack of context can create difficulties in distinguishing between implicit and explicit speech, sentences that are ableist but not anti-autistic, unrelated but not anti-autistic, and sentences discussing controversial entities in a neutral manner. For example:

> "[REDACTED] had a bad Autism Speaks race. I feel bad for them."

This sentence was difficult to label as the context in which the original poster is referring to Autism Speaks is unclear in this post, and the annotators are unfamiliar with the event being referenced. Autism Speaks defines itself as an organization raising "awareness" for autism [2] that autistic activists criticize for perpetuating dehumanizing stereotypes [77]. Therefore, this sentence may have various connotations depending on the nature of the event, the tone of the original poster, and other relevant context.

4.1.4 Differences in Annotator Labeling Strategies. We uncover the differences in annotator labeling strategies which contribute to their perceptions of anti-autistic ableism expressed in the sentences. These include:

(1) **Source Identification:** identifying the source of the text such as the name of the Sub-Reddit and whether or not the original poster is actually autistic. This may provide more context on the tone of the sentences, as sarcasm and other figurative speech may not be detectable in a single sentence.
(2) **Explicit Slurs:** looking for any explicit slurs and how they are used in the sentence to determine whether the sentence is anti-autistic or ableist.

[2]https://www.autismspeaks.org/

(3) **Overall Impact:** assessing the impact the sentence may have on the target group. Annotators look for the presence of violent or graphic language, and assess whether the sentence will cause any direct or indirect harm to a single person or the entire group.

This practical approach is employed by some annotators to determine whether a particular sentence should be classified as abelist speech and whether or not it is explicit by nature. For example, the following sentence has a high level of disagreement among annotators:

> "if they have never 'handled' an autistic kid before, don't feel bad! They probably did not know.. ?"

Annotators find it difficult to label as the original poster's identity, tone, and impact on autistic people were unclear. The quotation marks may suggest the original poster was being sarcastic, or is quoting someone. The missing context makes the sentence too ambiguous for a consensus on its classification.

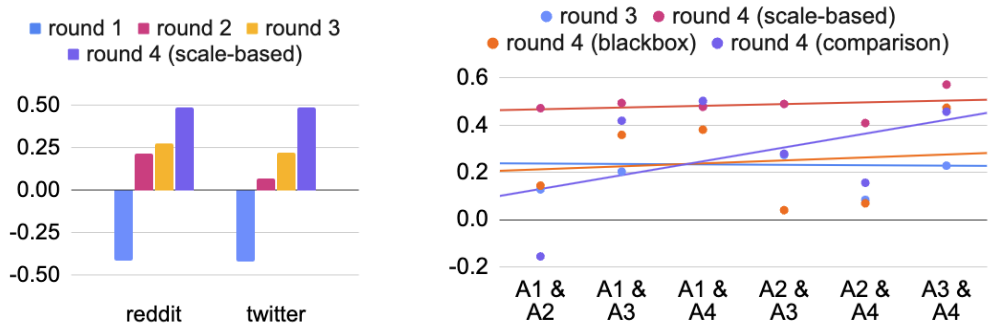
4.1.5 Individual vs. Collaborative Labeling Preferences. While annotators seem to prefer a granular approach individually, the results of our co-design session reveal that in group settings, a binary classification was preferred. Interestingly, the discussion among annotators reveals that they feel less confident labeling sentences as not ableist in a group setting, opting for the "unrelated" category instead. Although half of the annotators select "ableist but not anti-autistic" labels for certain sentences in their third round of annotations, in the group annotation, only the "explicitly anti-autistic" label is applied to these sentences, resulting in a binary classification of sentences as either "anti-autistic" or not. After completing the group annotation task, one annotator emphasizes the need to have an annotation scheme that allows participants to go back and alter their previous labels.

4.2 Agreement Scores

The inter-agreement scores for subjective tasks such as hate speech [51, 65, 80] or other complex tasks such as claim matching [39] tend to be low. Therefore, the starting low agreement scores, such as -.3 in round 1 and .2 in round 2, which can be observed in Figure 3, were expected. In figure 3, we share the improvement in our annotator the scores throughout each iteration from worse-than-chance to moderate agreement. While increasing the granularity of labels by separating unrelated sentences initially results in an improvement in agreement, we observe little improvement among rounds 2 and 3 when we separate ableist and anti-autistic speech. This indicates that making the annotation task more focused on a specific identity (e.g. anti-autistic speech instead of ableist speech), and less focused on the nature of the speech (e.g. as implicit or explicit) contributes to large improvements in annotator agreement. As well, it also improves agreement among our annotator pairs who tend to consistently have lower levels of agreement, as shown in Figure 3. During our co-design session, we observed that the comparison labeling task requires twice as much time from the annotators as scale-based annotation, as the annotators have to read through and process two different sentences and then compare them to each other, which further complicates the annotation task. Comparison labeling ultimately leads to the lowest agreement among our annotator pairs, while the binary scale consistently gives the highest overall agreement.

4.3 Limitations

We are building upon prior work demonstrating the harms of applying a deficits-based medical model understanding of autism in research [37, 61, 63, 71, 77]. However, this approach is largely centered on Western perspectives as disabilities including autism are defined in different ways around the world [40, 60]. Additionally, our usage of person-first language (i.e. "autistic people")



(a) (b)

Fig. 3. Comparing our annotator agreement using different labeling schemes (*a*) and annotation techniques (*b*). Although the granularity of the labels varies as discussed in Section 3.4, these scores have been converted to a binary classification of "anti-autistic" or not to allow for a fair comparison.

aligns with the perspectives of 87% of autistic adults in the United States who prefer it over identity-first language (i.e. "people with autism") [74]. We note these preferences also vary across cultures as prior work has shown there is no universally accepted way to talk about autism [40]. While our annotators are racially diverse, all of them are English-speaking Westerners and thus their perspectives may not be representative of international and intercultural perspectives. All of these aspects impact our ground-truth and other conclusions, and therefore our findings may not be generalizable across different cultures.

4.3.1 Considerations for Automated Content Moderation. Since the language used to describe neuro-divergence is constantly changing as cultural and social attitudes shift, future work should continue to explore the alignment of AI classifications with current community perspectives to ensure these systems are more accurate. For example, autistic individuals have reclaimed traditionally pejorative terms into expressions of identity and empowerment. This includes the label "autistic" itself, which was once utilized as an in-group insult or out-group slur [15]. However, these discussions may be unfairly censored by content moderation systems that are not trained to recognize positive intra-community discussions [54]. Similarly, these systems may miss discriminatory language that is not explicit by nature, or language that is considered offensive by the community, but is still commonly used by others [9]. The constant evolution of formal and informal language on social media platforms creates a continuous challenge for AI models to adapt and keep pace with these changes. This problem is further exacerbated with the use of alternative spellings or emojis to bypass content moderation filters [13]. Hence, there is a critical need for ensuring human-labeled datasets continuously and effectively incorporate community perspectives if they are used to create automated content moderation systems.

5 Discussion

The importance of examining and mitigating annotator biases have been investigated by prior works at CSCW [7, 87]. One of these works focuses on better integrating the perspectives of the impacted stakeholders, and introduces a dataset contextualization tool enabling AI practitioners to work toward more ethical systems [87]. While this work focuses on contextualizing outputs, our work contextualizes inputs by enabling annotators to view the data in context, and collaboratively

examine biases and other limitations. The iterative process with annotator discussions provide us with a clearer understanding of the difficulties of such tasks, and suggestions from annotators on addressing these difficulties to increase agreement. We expand on these in the following sections.

5.1 What challenges do annotators face that lead to disagreements when labeling anti-autistic language?

Disagreements among annotators, while often viewed negatively, are essential to the iterative process. They provide valuable insights into the varying perspectives and interpretations that annotators bring to the table [57]. Recognizing and addressing these disagreements are crucial, as they highlight the necessity for major revisions in annotation guidelines. The iterative process appears to be more effective in dealing with the fluid nature of subjective tasks and, in this case, results in improvements to our task guidelines.

Vague Definitions: Clearer definitions of terms like anti-autistic ableism are necessary, as people's understanding of and sensitivity toward recognizing such speech were added to the document. We initially use the following definitions for our labels:

> *Implicitly Ableist*: describing autism using medical terminology such as an "illness" or "disease", or focus on clinical applications such as "curing" or "diagnosing" autism
>
> *Explicitly Ableist*: using ableist slurs, dehumanizing autistic people by comparing them to non-human entities such as animals, using anti-autistic language that promotes negative stereotypes, or expressing negative emotions such as fear, disgust, or hatred toward autistic people.

We updated these definitions in round 4, reflecting the findings of our discussions. Our initial definitions are written under the assumption that ableist language encompasses anti-autistic language, and that the explicitness of the speech will matter to the annotators. However, our annotators struggle to differentiate between ableist and anti-autistic speech. Thus, in round 4, the aforementioned definition for the "implicitly ableist" label is applied to "implicitly anti-autistic" speech instead. Meanwhile, the implicitly ableist label's definition is amended to include making assumptions about disabled people that would not be made for able-bodied people, referring to disabilities as "inspiring", using condescending language such as saying people "suffer from autism", and infantilizing disabled people such as calling them "innocent" or "pure" [9].

Overly Complex Annotation Schemes: Eventually, due to disagreements in their perceptions of ableist, anti-autistic, and implicit or explicit speech, our annotators' co-design an annotation scheme that is binary and hones in on the target group's identity. By eliminating the categories defining the nature of the speech as implicit or explicit, the annotators reveal this aspect is not as important to them as we initially expected it to be. The annotators prefer clearly defining and identifying the target group, which they believe can be achieved with identity-specific labeling schemes.

Lack of Context: Our annotator discussions show that over half of the difficult-to-label sentences need more context. This highlights the need to understand how annotators determine context. It is particularly crucial when labeling implicitly ableist sentences, as figurative speech like sarcasm is hard to identify without knowing important details like the speaker's identity. Our findings indicate the importance of providing annotators with the resources they need to deduce this context, such as providing more information on the source of the data, or displaying the sentences in-context from the conversations they were extracted from.

Inadequate Resources: While the right labeling scheme can reduce disagreements to some extent, training and other resources are necessary to increase agreement further. Most disagreements arose from different perceptions of intent and the explicitness of speech. In particular, the impact

of speech played a larger role in assessing implicit ableist language, such as the medical model, for some annotators. This underscores the complexity of the labeling task and the need for continuous improvement in guidelines and training to ensure more consistent and accurate annotations.

5.1.1 Annotator Perspectives on Addressing These Challenges: While prior research has demonstrated that collaborative decisions improve decision accuracy [38], there is a gap in research exploring annotator perspectives and recommendations, as other works at CSCW have focused primarily on the outputs of AI models [7, 33, 70, 83, 87], while we examine the dataset creation process itself.

Our findings contribute to research on mitigating biases and improving participatory AI by empirically showing that iterative and collaborative processes impact several aspects of the annotation process and the resulting ground truth data [7, 87]. When observing the annotators discuss and re-label sentences, we take notes on the different issues that are raised, particularly related to our labels and resources provided, such as the clarity of our annotation guidelines. Through our annotation process, we uncover the details of the challenges annotators face while classifying sentences and the ways in which they can be addressed. These findings highlight the importance of applying a more collaborative approach to subjective classification tasks such as ableist language detection.

Improving Resources: Providing training or orientation sessions can help annotators feel more confident in their labeling decisions, especially in complex scenarios where the distinction between ableist and non-ableist speech is not clear. However, other revisions, such as a glossary of key terms, can help provide more context to diverse annotators such as non-native English speakers or those unfamiliar with the medical terminology used in autism discourse by providing clearer and more comprehensive definitions.

Capturing Changes in Perspectives: Another significant outcome of the iterative process is the annotators' request to change labels after discussing past annotations with newer knowledge. For example, an annotator experienced a change in their perception of the following sentence:

> "Atypical response to the expression of fear and limited social orienting, joint attention, and attention to another's distress have also been reported in young children with autism"

While the annotators initially label this sentence as implicitly ableist, following the discussions, they label it as explicitly ableist. This was due to a change in the annotators' perceptions of words such as "atypical" being used to describe autistic people, as they gain a better understanding of neuronormativity or the belief that there is a "normal" brain that autistic people deviate from [84].

This highlights the dynamic nature of the annotation process and the need to rethink the implementation of models, especially for subjective tasks. The assumption that a constant model represents ground truth is challenged by the realization that people's opinions and interpretations can change over time. The language used to describe autism, in particular, is controversial due to its ties to Nazi eugenics or the broader ableism in our society [20]. While functioning labels such as "high-functioning' or "low-functioning" may still be used to separate autistic people based on their economic worthiness, these classifications are rooted in eugenicist beliefs [20]. Further, researchers have found preferences for person-first language, i.e. saying person with autism, are influenced by disability stigma [20]. As we increase our awareness and understanding of the harms of such speech, it is important for the technologies that we use to reflect these values [36].

Simplifying Labels: Furthermore, our findings contribute to prior CSCW research on mitigating biases and improving participatory AI by empirically showing that iterative and collaborative processes impact several aspects of the annotation process and the resulting ground truth data [7, 87].

The inclusion of annotator discussion significantly influenced the renaming of the categories (e.g. from 'ableist' to 'anti-autistic') and the categorization of labels from granular to binary.

While completing annotations individually, the annotators express interest in a category for 'unrelated' sentences. An example of a sentence which could be classified under this category is:

> "Arse, forgot about a webinar this morning. Now I'll never know how to secure virtualised environments"

Following the discussions in round 1, the annotators ask for a separate label for sentences such as this that are not related to autism or disability. However, during the group labeling session in round 4, they agree that the 'unrelated' and 'not ableist' categories can be consolidated together in a binary classification since such sentences do not focus on autism and therefore can not be anti-autistic.

5.2 What impact will this design process have on the perspectives of the annotators and dataset creators toward this annotation task?

Both the annotators' and dataset creators recognized the importance of testing different annotation strategies due to a misalignment with their perceptions and the outcomes of the task iterations. The changes in the annotators' preferences for labels before and after our tests are reflective of their understanding of the complexity of classifying anti-autistic langauge.

5.2.1 Annotator Perspectives. Our findings reveal a collaborative annotation process helps improve annotators' understanding of different perspectives. Through this, our annotators propose an approach to simplify the labels while encompassing different perspectives.

Bridging Perspectives: Through group discussions and the co-design session, the annotators gain a better understanding of how their peers approach the task. For example, some annotators try to determine the "impact" the sentence may have on an autistic person, i.e. if the sentence is promoting violence or harmful stereotypes. However, others focus more on trying to determine the original posters' identity or other situational context (e.g. the subreddit it was posted in) and the tone of the sentence in the classification task. These discussions help both the annotators and dataset creators gain insights on the disagreements over the classification of "implicit" and "explicit" hate speech, and the difficulty of defining these labels in a manner that can serve as a bridge between different individual perspectives and improve agreement.

We find that simply providing the annotators with guidelines and examples are not sufficient enough as they do not account for the different perspectives of every individual. Our findings echo the results of prior research as we find group discussions were necessary to help the annotators gain a better understanding of their peers' perspectives on the task [38, 58]. The discussions also encourage annotators to reflect on their personal challenges, which may not be evident to them in one-on-one discussions. Becoming aware of the perspectives of their peers helps them identify the similarities and differences that contribute to disagreements and makes them aware of their own biases. In other words, removing the annotators' isolation may help improve their annotation as it allows them an opportunity to reflect on a deeper level.

Simplifying Labels: In the early stages of the annotation process, the annotators favor a more granular approach toward identity-based classification (e.g. separate labels for 'anti-autistic' and 'ableist' speech). However, there are frequent disagreements among the annotation team and even the dataset creators on how to separate the two categories. For example, certain slurs such as r*tard are not exclusively used for autistic people, and may target other groups of people such as those with intellectual disabilities, which can make it difficult to identify the target group. However, even if such sentences may not target autistic people directly, they can still be considered anti-autistic as the slur is also used against autistic people [8]. Separating ableist speech based on identity adds

another layer of difficulty, especially if such speech intersects and harms multiple communities. During the co-design session, the annotators believe narrowing the scope to a single identity (e.g. only the 'anti-autistic' label), would simplify the annotation task.

5.2.2 Researcher Perspectives. Through this work, we improve our understanding of the task including how our perspectives as dataset creators differ from the annotators' perspectives, ways to improve the process, and other ethical considerations arising from converting a highly subjective task into an objective classification.

Implicit vs. Explicit Speech: While we initially assumed the distinction between implicit and explicit ableist speech would be important to annotators, our discussions reveal that disagreements arise based on their perceptions of the impact of the speech. For example, some annotators believe referring to autistic people as 'atypical' is explicitly anti-autistic as it defines neurotypicals as the norm and autistic people as deviating from that norm, in alignment with neuronormativity [84]. However, other annotators believe that such speech should be considered implicitly anti-autistic as it does not contain any slurs. This collaborative process encourages reflection on whether the distinction between implicit and explicit speech is important to preserve as the annotators display a clear preference for its removal.

Considering Impact: The dataset creators also reflect on how the annotators' varying approaches to the task can impact the outcome of their results. For example, some annotators focus on the 'impact' of the speech, which they define as how harmful it will be in the immediate future for the autistic community. This approach has its limitations as for implicit hate speech, the harms may not be as evident. For example, a sentence such as 'vaccines will give your child autism' may not immediately appear to be harmful, but it promotes misrepresentations and stereotypes which contribute to the broader marginalization of autistic people in our society.

Creating Flexible Schemes: We also gain deeper insights on the dynamic nature of this task and how it creates difficulties in the annotation process. We learn the importance of creating an annotation scheme that is flexible, as annotators express an interest in being able to go back and re-label certain sentences based on newly gained knowledge. This can be implemented in a variety of ways. For example, for human-annotated datasets such as ours, we can introduce an edit feature in the annotation script that allows annotators to update their labels. Further, other researchers can also implement this through methods such as in-context learning for LLMs [21]. While we anticipated such changes due to our ever-evolving understanding and acceptance of autism in society, we did not expect annotators may be interested in going back and editing their labels after they had already completed the annotation task.

Avoiding Over-Simplification: We reflect on the trade-off between simplifying the annotation task while preserving community perspectives. Oversimplifying the task may impact the accuracy of the classification. In a purely granular scheme, sentences that need more context will be classified as 'not anti-autistic' because some annotators may be unable to identify the tone of the sentences, even if it appears to be implicitly ableist. Similarly, the reclamation of slurs by community members may be inaccurately classified as 'anti-autistic' if the annotator cannot determine the speaker's identity. This can contribute to unfair censorship if such a classification is used for content moderation in the real world. This is similar to what happens in hate speech for terms reclaimed by other communities [66].

Improving Resources: These discussions help the dataset creators gain insights on how the tools we provide to annotators, such as term definitions and annotation guidelines, may be interpreted in various ways by a diverse group of annotators. Ultimately, these discussions help us obtain annotator feedback that is useful in refining the task description, resources we provide to annotators, and the task itself. We find that a collaborative approach will improve the annotator and dataset

creators' understanding of the task. For example, our annotators indicated that the definitions and examples we included in our annotation guidelines were not enough as they were not exhaustive and subject to differences in interpretations. Thus, we introduce a glossary (see Appendix B) in our annotator guidelines in round 4 as a dynamic resource that annotators can update as needed to define any new terms and their connotations as they encounter them in the annotation process. This glossary is viewed favorably by annotators as the words they identified during the annotation processes are included. This collaborative resource is different from the initial definitions and example sentences we provided, which we wrote and edited and could not be updated during the annotation process.

Ethical Considerations: Finally, these discussions lead us to reflect on the ethical considerations of trying to quantify a dynamic and multi-faceted phenomenon that is largely influenced by personal biases using a single metric, or in this case, relying on annotator agreements to determine 'accuracy'. Ultimately, the biggest limitation of this task for the dataset creators is the difficulty of converting a subjective task into an objective classification. One of the main limitations of automating subjective tasks such as classifying anti-autistic hate speech is defining our ground truth. While the annotators' perspectives are limited by their own identities and experiences [18], our work as a whole is also limited by how we define anti-autistic ableism, and how we handle disagreements. For instance, if we have a majority of non-autistic annotators who are not as sensitive to recognizing anti-autistic speech as our minority autistic annotators. In such a scenario, our work will inherently center the perspectives of non-autistic people especially if we use a majority-wins approach when handling disagreements.

In recent years, researchers have tried to incorporate minority opinions in their ground truth data in efforts to provide a more granular overview of community perspectives as the majority opinion may not accurately represent the diversity of opinions [82]. However, it is important to note that the way we define and understand disabilities, including autism, is ever-evolving, and varies both on an individual and systemic level by identity, culture, language, and other factors [6, 60]. Thus, unless the classification is personalizable and dynamically updates as needed, it is difficult to truly generate a universal 'ground truth'.

Future work should include the diverse perspectives of autistic people to find the right balance between the community's perspective and the annotators'. Such work can determine what aspects of anti-autistic speech are important to them so those aspects can be preserved in the data annotation process. For example, such work can explore whether or not the distinction between implicit and explicit anti-autistic speech may be more important to autistic people than it is to the annotators, which can help dataset creators determine the most effective labeling scheme for their work.

6 Conclusion

We adopt annotator-centric and collaborative methods to examine how annotators approach anti-autistic speech annotations. Our work contributes to CSCW research focusing on investigating and mitigating annotator biases, adapting participatory AI methodologies, and testing novel tools and techniques for human annotation [7, 44, 87]. We address a gap in research focusing on the dataset creation process itself, as prior research at CSCW has focused mainly on biases in model outputs [7, 33, 70, 83, 87].

Our findings indicate that collaborative annotation method can help improve the annotators' understanding of their peers' perspectives and their own biases, thus improving their coordination and agreement on the annotation task. Additionally, we explore the effectiveness of improving the resources provided to annotators in clarifying and simplifying the task for them. Future CSCW research can build upon these insights when creating novel annotation tools. For example, providing a glossary of commonly used terms and their definitions that can be collaboratively updated by

the annotators, and allowing them to re-label certain sentences are useful adaptations to support diverse annotator teams in labeling ever-evolving language.

We also examine the practical limitations of over-simplifying the annotation task, as it may lead to community censorship or neglecting hateful content when used in automated content moderation systems. Therefore, future work must examine the trade-off between preserving community perspectives while simplifying the task for annotators to determine what aspects of the labels to preserve. This includes studying the alignment of the sentence classification with the perspectives of autistic people. For example, in examining whether or not the distinction between explicit and implicit speech matters in this task to autistic people.

References

[1] Basma Alharbi, Hind Alamro, Manal Alshehri, Zuhair Khayyat, Manal Kalkatawi, Inji Ibrahim Jaber, and Xiangliang Zhang. 2020. ASAD: A twitter-based benchmark arabic sentiment analysis dataset. *arXiv preprint arXiv:2011.00578* (2020).

[2] Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information* 13, 6 (2022), 273.

[3] DSMTF American Psychiatric Association, American Psychiatric Association, et al. 2013. *Diagnostic and Statistical Manual of mental disorders: DSM-5*. Vol. 5. American Psychiatric Association Washington, DC.

[4] Valerio Basile et al. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *CEUR workshop proceedings*, Vol. 2776. CEUR-WS, 31–40.

[5] Herbert Bless and Klaus Fiedler. 2014. *Social cognition: How individuals construct social reality*. Psychology Press.

[6] Kathleen R Bogart, Samuel W Logan, Christina Hospodar, and Erica Woekel. 2019. Disability models and attitudes among college students with and without disabilities. *Stigma and Health* 4, 3 (2019), 260.

[7] Nattapat Boonprakong, Gaole He, Ujwal Gadiraju, Niels van Berkel, Danding Wang, Si Chen, Jiqun Liu, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. 2023. Workshop on Understanding and Mitigating Cognitive Biases in Human-AI Collaboration. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 512–517. doi:10.1145/3584931.3611284

[8] Monique Botha, Bridget Dibb, and David M Frost. 2022. " Autism is me": an investigation of how autistic individuals make sense of autism and stigma. *Disability & Society* 37, 3 (2022), 427–453.

[9] Kristen Bottema-Beutel, Steven K Kapp, Jessica Nina Lester, Noah J Sasson, and Brittany N Hand. 2021. Avoiding ableist language: Suggestions for autism researchers. *Autism in adulthood* (2021).

[10] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[11] Simon M Bury, Rachel Jellett, Alex Haschek, Michael Wenzel, Darren Hedley, and Jennifer R Spoor. 2023. Understanding language preference: Autism knowledge, experience of stigma and autism identity. *Autism* 27, 6 (2023), 1588–1600.

[12] Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 6 (Jun. 2023), 6860–6868. doi:10.1609/aaai.v37i6.25840

[13] Kendra Calhoun and Alexia Fawcett. 2023. "They Edited Out her Nip Nops": Linguistic Innovation as Textual Censorship Avoidance on TikTok. *Language@internet* 21 (2023), 1–.

[14] Benedetta Catanzariti, Srravya Chandhiramowuli, Suha Mohamed, Sarayu Natarajan, Shantanu Prabhat, Noopur Raval, Alex S. Taylor, and Ding Wang. 2021. The Global Labours of AI and Data Intensive Systems. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 319–322. doi:10.1145/3462204.3481725

[15] Bianca Cepollaro, Marta Jorba, and Valentina Petrolini. [n. d.]. The case of 'Autistic'. Pejorative Uses and Reclamation. *Ergo. An Open Access Journal in Philosophy* ([n. d.]).

[16] Srravya Chandhiramowuli. 2024. Making AI Work: Tracing Human Labour in the Supply Chains of Dataset Production. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing* (San Jose, Costa Rica) *(CSCW Companion '24)*. Association for Computing Machinery, New York, NY, USA, 47–49. doi:10.1145/3678884.3682055

[17] Elizabeth Chase. 2017. Enhanced member checks: Reflections and insights from a participantresearcher collaboration. *The Qualitative Report* 22, 10 (2017), 2689–2703.

[18] Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics* 11 (2023), 300–319.

doi:10.1162/tacl_a_00550
[19] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516* (2019).
[20] Anna N De Hooge. 2019. Binary boys: autism, aspie supremacy and post/humanist normativity. *Disability Studies Quarterly* 39, 1 (2019).
[21] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234* (2022).
[22] Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the Majority is Wrong: Modeling Annotator Disagreement for Subjective Tasks. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:258615411
[23] Carli Friedman. [n. d.]. Ableist Language & Disability Professionals: Differences in Language Based on Professionals' Demographics. ([n. d.]).
[24] Maysa Gama. [n. d.]. Artificially Intelligent and Implicitly Ableist: Analyzing Language Model Bias Towards Persons with Disabilities. ([n. d.]).
[25] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (2020), 2053951720943234.
[26] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford* 1, 12 (2009), 2009.
[27] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862* (2019).
[28] John Greally. 2021. AUTISM GLOSSARY: ACRONYMS & ABBREVIATIONS — linkedin.com. https://www.linkedin.com/pulse/autism-glossary-acronyms-abbreviations-john-greally/. [Accessed 19-06-2024].
[29] Salim Hafid, Sebastian Schellhammer, Sandra Bringay, Konstantin Todorov, and Stefan Dietze. 2022. SciTweets-A Dataset and Annotation Framework for Detecting Scientific Online Discourse. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 3988–3992.
[30] Matan Halevy, Camille Harris, Amy Bruckman, Diyi Yang, and Ayanna Howard. 2021. Mitigating Racial Biases in Toxic Language Detection with an Equity-Based Ensemble Framework. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (–, NY, USA) *(EAAMO '21)*. Association for Computing Machinery, New York, NY, USA, Article 7, 11 pages. doi:10.1145/3465416.3483299
[31] Melinda C Hall. 2019. Critical disability theory. (2019).
[32] Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures* 1, 1 (2007), 77–89.
[33] Danula Hettiachchi, Mark Sanderson, Jorge Goncalves, Simo Hosio, Gabriella Kazai, Matthew Lease, Mike Schaekermann, and Emine Yilmaz. 2021. Investigating and Mitigating Biases in Crowdsourced Data. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 331–334. doi:10.1145/3462204.3481729
[34] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813* (2020).
[35] Shivani Kapania, Alex S Taylor, and Ding Wang. 2023. A hunt for the snark: Annotator diversity in data practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.
[36] Steven K Kapp. 2023. Profound concerns about "profound autism": Dangers of severity scales and functioning labels for support needs. *Education Sciences* 13, 2 (2023), 106.
[37] Steven K Kapp, Kristen Gillespie-Lynch, Lauren E Sherman, and Ted Hutman. 2013. Deficit, difference, or both? Autism and neurodiversity. *Developmental psychology* 49, 1 (2013), 59.
[38] Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2022. What makes you change your mind? An empirical investigation in online group decision-making conversations. arXiv:2207.12035 [cs.CL] https://arxiv.org/abs/2207.12035
[39] Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott Hale. 2021. Claim Matching Beyond English to Scale Global Fact-Checking. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 4504–4517. doi:10.18653/v1/2021.acl-long.347
[40] Connor Tom Keating, Lydia Hickman, Joan Leung, Ruth Monk, Alicia Montgomery, Hannah Heath, and Sophie Sowden. 2023. Autism-related language preferences of English-speaking individuals across the globe: A mixed methods investigation. *Autism Research* 16, 2 (2023), 406–428.
[41] Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality Management in the Wild. *Computational Linguistics* 50, 3 (2024), 817–866.

[42] Sami Koivunen, Saara Ala-Luopa, Thomas Olsson, and Arja Haapakorpi. 2022. The march of Chatbots into recruitment: recruiters' experiences, expectations, and design opportunities. *Computer Supported Cooperative Work (CSCW)* 31, 3 (2022), 487–516.

[43] Emma Lagren. 2023. *Artificial intelligence as a tool in social media content moderation.* B.S. thesis.

[44] Anna De Liddo, Ágnes Sándor, and Simon Buckingham Shum. 2011. Contested Collective Intelligence: Rationale, Technologies, and a Human-Machine Annotation Study. *Computer Supported Cooperative Work (CSCW)* 21 (2011), 417 – 448. https://api.semanticscholar.org/CorpusID:6927981

[45] Vittorio Lingiardi, Nicola Carone, Giovanni Semeraro, Cataldo Musto, Marilisa D'Amico, and Silvia Brena. 2020. Mapping Twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis. *Behaviour & Information Technology* 39, 7 (2020), 711–721.

[46] Emma Llansó, Joris VaN hoboKeN, Paddy Leerssen, and Jaron Harambam. 2020. Content Moderation, and Freedom of Expression. *Algorithms* (2020).

[47] Marta Marchiori Manerba and Sara Tonelli. 2021. Fine-grained fairness analysis of abusive language detection systems with checklist. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. 81–91.

[48] Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2023. Automated Ableism: An Exploration of Explicit Disability Biases in Sentiment and Toxicity Analysis Models. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, Anaelia Ovalle, Kai-Wei Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop Kumar, and Rahul Gupta (Eds.). Association for Computational Linguistics, Toronto, Canada, 26–34. doi:10.18653/v1/2023.trustnlp-1.3

[49] Selin E Nugent and Susan Scott-Parker. 2022. Recruitment AI has a Disability Problem: anticipating and mitigating unfair automated hiring decisions. In *Towards trustworthy artificial intelligent systems*. Springer, 85–96.

[50] World Health Organization. [n. d.]. Autism. https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders. [Accessed 28-06-2024].

[51] 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 4675–4684. doi:10.18653/v1/D19-1474

[52] Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing Toxic Content in Large Pre-Trained Language 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 4262–4274. doi:10.18653/v1/2021.acl-long.329

[53] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 2799–2804. doi:10.18653/v1/D18-1302

[54] Sarah M Parsloe. 2015. Discourses of disability, narratives of community: Reclaiming an autistic identity online. *Journal of Applied Communication Research* 43, 3 (2015), 336–356.

[55] Kamakshya P Patra and Orlando De Jesus. 2020. Echolalia. (2020).

[56] Desmond Patton, Philipp Blandfort, William Frey, Michael Gaskell, and Svebor Karaman. 2019. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. (2019).

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

[58] Dave Randall. 2016. What Is Common in Accounts of Common Ground? *Comput. Supported Coop. Work* 25, 4–5 (oct 2016), 409–423. doi:10.1007/s10606-016-9256-7

[59] Gerald Rau and Yu-Shan Shih. 2021. Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *Journal of english for academic purposes* 53 (2021), 101026.

[60] Marno Retief and Rantoa Letšosa. 2018. Models of disability: A brief overview. *HTS Teologiese Studies/Theological Studies* 74, 1 (2018).

[61] Larry Arnold Richard Woods, Damian Milton and Steve Graby. 2018. Redefining Critical Autism Studies: a more inclusive interpretation. *Disability & Society* 33, 6 (2018), 974–979. arXiv:https://doi.org/10.1080/09687599.2018.1454380 doi:10.1080/09687599.2018.1454380

[62] Naba Rizvi, Andrew Begel, and Hala Annabi. 2021. Inclusive Interpersonal Communication Education for Technology Professionals.. In *AMCIS*.

[63] Naba Rizvi, William Wu, Mya Bolds, Raunak Mondal, Andrew Begel, and Imani NS Munyaka. 2024. Are Robots Ready to Deliver Autism Inclusion?: A Critical Review. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[64] Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why Don't You Do It Right? Analysing Annotators' Disagreement in Subjective Tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 2428–2441. doi:10.18653/v1/2023.eacl-main.178

[65] 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)*, Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1443

[66] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 1668–1678.

[67] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997* (2021).

[68] Amy Sequenzia. 2014. Is Autism Speaks a Hate Group? - Autistic Women & Nonbinary Network (AWN) — awnnetwork.org. https://awnnetwork.org/is-autism-speaks-a-hate-group/. [Accessed 19-06-2024].

[69] Jonathan A Smith, Michael Larkin, and Paul Flowers. 2021. Interpretative phenomenological analysis: Theory, method and research. (2021).

[70] Jaemarie Solyst, Ellia Yang, Shixian Xie, Amy Ogan, Jessica Hammer, and Motahhare Eslami. 2023. The Potential of Diverse Youth as Stakeholders in Identifying and Mitigating Algorithmic Bias for a Future of Fairer AI. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 364 (Oct. 2023), 27 pages. doi:10.1145/3610213

[71] Katta Spiel, Christopher Frauenberger, Os Keyes, and Geraldine Fitzpatrick. 2019. Agency of autistic children in technology research—A critical literature review. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 6 (2019), 1–40.

[72] Katta Spiel and Kathrin Gerling. 2021. The purpose of play: How HCI games research fails neurodivergent populations. *ACM Transactions on Computer-Human Interaction (TOCHI)* 28, 2 (2021), 1–40.

[73] Katta Spiel, Eva Hornecker, Rua Mae Williams, and Judith Good. 2022. ADHD and technology research–investigated by neurodivergent readers. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–21.

[74] Amanda Taboas, Karla Doepke, and Corinne Zimmerman. 2023. Preferences for identity-first versus person-first language in a US sample of autism stakeholders. *Autism* 27, 2 (2023), 565–570.

[75] Zeerak Talat. 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*, David Bamman, A. Seza Doğruöz, Jacob Eisenstein, Dirk Hovy, David Jurgens, Brendan O'Connor, Alice Oh, Oren Tsur, and Svitlana Volkova (Eds.). Association for Computational Linguistics, Austin, Texas, 138–142. doi:10.18653/v1/W16-5618

[76] Kaggle Team. 2019. May 2015 Reddit Comments. https://www.kaggle.com/datasets/kaggle/reddit-comments-may-2015 [Accessed 20-06-2024].

[77] Ronnie Thibault. 2014. Can autistics redefine autism? The cultural politics of autistic activism. *Trans-scripts* 4 (2014), 57–88.

[78] Shane Timmons, Frances McGinnity, and Eamonn Carroll. 2024. Ableism differs by disability, gender and social context: Evidence from vignette experiments. *British Journal of Social Psychology* 63, 2 (2024), 637–657.

[79] Pranav Narayanan Venkit and Shomir Wilson. 2021. Identification of bias against people with disabilities in sentiment analysis and toxicity detection models. *arXiv preprint arXiv:2111.13259* (2021).

[80] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In *Italian Conference on Cybersecurity*. https://api.semanticscholar.org/CorpusID:8293149

[81] Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 14523–14530.

[82] Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023. Disagreement Matters: Preserving Label Diversity by Jointly Modeling Item and Annotator Label Distributions with DisCo. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4679–4695. doi:10.18653/v1/2023.findings-acl.287

[83] Lauren Wilcox, Robin Brewer, and Fernando Diaz. 2023. AI Consent Futures: A Case Study on Voice Data Collection with Clinicians. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 316 (Oct. 2023), 30 pages. doi:10.1145/3610107
[84] Sonny Jane Wise. 2023. *We're All Neurodiverse: How to Build a Neurodiversity-Affirming Future and Challenge Neuronormativity.* Jessica Kingsley Publishers.
[85] Michael Yoder, Lynnette Ng, David West Brown, and Kathleen Carley. 2022. How Hate Speech Varies by Target Identity: A Computational Analysis. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, Antske Fokkens and Vivek Srikumar (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 27–39. doi:10.18653/v1/2022.conll-1.3
[86] Bingjie Yu, Katta Spiel, and Leon Watts. 2019. Caring About Dissent: Online Community Moderation and Norm Evolution. In *Volunteer Work: Mapping the Future of Moderation Research - A CSCW '19 Workshop.*
[87] Angie Zhang. 2024. Empowering and Centering Impacted Stakeholders in AI Design. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing* (San Jose, Costa Rica) *(CSCW Companion '24)*. Association for Computing Machinery, New York, NY, USA, 50–53. doi:10.1145/3678884.3682053
[88] Annuska Zolyomi and Jaime Snyder. 2021. Social-emotional-sensory design map for affective computing informed by neurodivergent experiences. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–37.
[89] Âutistic ûnion. 2012. What does the Âû after my name mean? https://www.facebook.com/AutisticUnion/posts/456270757745263/. [Accessed 19-06-2024].

A Guidelines Provided Overtime

A.1 Sentence-Level Annotation

We are annotating text extracted from human-robot interaction literature and social media (Reddit, Twitter). The goal of this study is to classify sentences as either i) implicitly ableist or ii) explicitly ableist iii) implicitly anti-autistic, iv) explicitly anti-autistic, v) not ableist, or vi) needs more context and/or is irrelevant. You will see paragraphs and sentences extracted from scientific literature focusing on robotics for autism published in conferences and journals, and interactions between anonymous social media users. These sentences may or may not contain ableist content.

A.1.1 Defining Ableism According to the Center for Disability Rights (emphasis added by us): Ableism is a set of beliefs or practices that devalue and discriminate against people with physical, intellectual, or psychiatric disabilities and often rests on the assumption that disabled people need to be 'fixed' in one form or the other.

A.1.2 Neurodiversity vs. Neuroableism. We center neurodiversity in our work, which means we acknowledge that humans have different ways of thinking, communicating, and experiencing the world, and that these differences are valid forms of human diversity.

The belief that autistic people are "deficient" in certain skills is rooted in eugenicism and dehumanizing research originating in Nazi Germany, and contradict neurodiversity by promoting neuroableism, or the belief that neurotypical people are "normal", and people with various neurological differences such as autism, ADHD, or dyslexia are "abnormal". Such understandings may result in researchers studying autism as a "disorder" that needs to be diagnosed, treated, prevented, or cured. This work is ableist because it expects autistic people to be "fixed" by changing their communicational styles and behaviors to adapt to neurotypical social norms.

A.1.3 Not-Ableist (Round 1-3). These sentences may be entirely unrelated to autism or disabilities, or be written by an autistic person reaching out for help and support.

(1) Using medical terminology in a more personal manner (e.g. 'I need therapy').
(2) Discussions + suggestions from neurodivergent people (community-generated discussions).
(3) General discussions of the medical processes (unrelated to neurodivergence).
(4) Example: "I am autistic", "As an autistic person, I think . . . "

A.1.4 Implicitly Ableist (Round 1-3). Critical disability theory (CDT) is a framework centering disability which challenges the ableist assumptions present in our society. Using this theory, we define "implicitly ableist" content as:

(1) Making assumptions about a disabled person's abilities that would not be made about an able-bodied person.
(2) "Inspiration porn" → looking at disabled people as "inspiration" for able-bodied people
(3) Using condescending language such as saying a disabled person "suffers" with their disability
(4) Infantilizing disabled people by portraying them as unable to make their own decisions, or "innocent" and "pure"
(5) Example: "She is confined to a wheelchair", "it must be awful living with bipolar disorder"

A.1.5 Implicitly Anti-Autistic (Round 3). For the purposes of this study, we are applying a critical disability framework and classifying any sentences that describe autism using medical terminology such as an "illness" or "disease", or focus on clinical applications such as "curing" or "diagnosing" autism as implicitly ableist. These works are considered "implicitly ableist" as they do not use explicitly violent language or slurs, but still may be considered offensive according to critical autism theory.

- Using medical terminology to describe autism (e.g. saying its a "disorder") E.g., referring to autism as a public health crisis.
- Using language that may promote the infantilization or pathologization of autistic people
- Using words like "typically developing" to refer to non-autistic people.
- Saying autistic people have 'atypical behavior'.
- Ableist jokes using medical terminology (e.g. "It cured every disease, but was it's own cancer.")
- Other anti-autistic language (from the Bottema-Beutel paper) Example: "I am so proud of my autistic son for baking this cake", "even though she is autistic, she is a very good salesperson"

A.1.6 Explicitly Ableist (Round 1-3). Sentences using slurs for disabled people such as: retard, lame, insane, deluded, moron

A.1.7 Explicitly Anti-Autistic (Round 3). This category includes sentences that dehumanize autistic people (e.g. by comparing them to animals or non-living things), containing ableist slurs such as the r-word, other anti-autistic language that promotes negative stereotypes, or expressing negative emotions such as fear, disgust, or hatred toward autistic people. Slurs like the r-word specifically aimed at autistic people: Crazy, Delusional, Deranged, Dumb, Handicap, Idiot, Insane, Maniac, Moron, Madness, Stupid, Wheelchair bound, Differently abled, Handicapable, People with abilities, Special needs (source: [3]). Saying phrases like "autism intensifies" Using "autistic" as an insult (e.g. "that's so autistic")

A.1.8 Unrelated/Needs More Context (Round 1-4). This category includes text that is completely unrelated to disabilities, needs more context, and/or contains forms of media other than text. Examples include: tweets with images, ACM CCS categories, etc. Examples:

Not a sentence: "Keywords— Socially assistive robots; Affective robots; ASD; Autism; Developmental ability; Mullen; ADOS"

Includes non-text media: "@RonTerrell http://twitpic.com/3iqv9 - I wish I was there it looks like fun."

Needs more context: "You mean a walk won't cure an abscess in her mouth?"

[3]https://www.c-q-l.org/resources/articles/ableist-language-disability-professionals-differences-in-language-based-on-professionals-demographics/

B Glossary

An excerpt of our glossary developed as a result of annotators feedback is shown below in Table 1.

Received July 2024; revised December 2024; accepted March 2025

Table 1. Glossary of Terms

Term	Definition
Autism Speaks	Controversial group promoting autism 'awareness', some autistic people it should be considered a hate group
ASD	Autism Spectrum Disorder
DSM-V	The Diagnostic and Statistical Manual of Mental Disorders, 5th Edition published by the American Psychiatric Association (used to diagnose autism)
AuDHD	Having a combination of autism and ADHD
Au, Âû	Used by autistic individuals to self-identify as autistic
ND	Neurodivergent or neurodiverse
NT	Neurotypical
Allistic	Non-Autistic
Aspie	Someone with Aspergers (outdated term, considered offensive unless it is being used to self-identify)
Autie	Autistic
Autism mom, dad, or parent	A parent or caregiver of an autistic individual
Dx	Diagnose
Martyr mom or parent	A parent using their child's autism to gain sympathy, attention, etc.
self-Dx	Self-diagnose
OT	Occupational therapist
AA	Actually autistic
ABA	Applied Behavioural Analysis, controversial treatment for autism that has been linked to PTSD in autistic individuals
Masking	Suppressing behaviors such as stimming to appear neurotypical
SPD	Sensory processing disorder
Stimming, stim	Repetitive actions such as hand-flapping and singing used for self-soothing
Functioning label	Outdated and offensive way to describe the 'severity' of someone's autism
Echolalia	Repeating sounds, words, or phrases, often unintentionally
Co-morbid	Having more than one medical condition together
CW/TW	Content warning, trigger warning tags. Often used in discussions where the content may be graphic or sensitive in nature
ED	Executive dysfunction, such as challenges with planning, time management, organization, and completing tasks often experienced by ND individuals
IFL/PFL	Identity-first language or people-first language. Autistic individuals may prefer IFL (i.e. saying 'autistic person' instead of 'person with autism'
RSD	**Rejection sensitivity dysphoria, sensitivity to rejection often experienced by ND individuals**
2e	Twice exceptional: an ND individual who is also considered 'gifted'
AAC	Augmentative & Alternative Communication: ways to communicate besides speaking