How Crowd Worker Factors Influence Subjective Annotations: A Study of Tagging Misogynistic Hate Speech in Tweets

Danula Hettiachchi,[1] Indigo Holcombe-James,[1] Stephanie Livingstone,[1] Anjalee de Silva,[2] Matthew Lease,[3] Flora D. Salim,[4] Mark Sanderson [1]

[1] RMIT University, Australia, [2] The University of Melbourne, Australia,
[3] University of Texas at Austin, United States, [4] University of New South Wales, Australia
danula.hettiachchi@rmit.edu.au, indigo.holcombe-james@rmit.edu.au, stephanie.livingstone@student.rmit.edu.au, adesilva@unimelb.edu.au, ml@utexas.edu, flora.salim@unsw.edu.au, mark.sanderson@rmit.edu.au

Abstract

Crowdsourced annotation is vital to both collecting labelled data to train and test automated content moderation systems and to support human-in-the-loop review of system decisions. However, annotation tasks such as judging hate speech are subjective and thus highly sensitive to biases stemming from annotator beliefs, characteristics and demographics. We conduct two crowdsourcing studies on Mechanical Turk to examine annotator bias in labelling sexist and misogynistic hate speech. Results from 109 annotators show that annotator political inclination, moral integrity, personality traits, and sexist attitudes significantly impact annotation accuracy and the tendency to tag content as hate speech. In addition, semi-structured interviews with nine crowd workers provide further insights regarding the influence of subjectivity on annotations. In exploring how workers interpret a task — shaped by complex negotiations between platform structures, task instructions, subjective motivations, and external contextual factors — we see annotations not only impacted by worker factors but also simultaneously shaped by the structures under which they labour.

Introduction

Social media platforms allow users to create and share content with one another at an unprecedented scale. While this provides an incredibly powerful channel for human communication, the lack of accountability and the possibility to remain anonymous on online platforms have resulted in various online safety concerns, such as toxic language in various forms (Fortuna and Nunes 2018; MacAvaney et al. 2019; Poletto et al. 2021; Schmidt and Wiegand 2017; Vidgen and Derczynski 2020) (*e.g.,* hate speech, abusive language, offensive language, etc.). Robust and reliable automated tooling, coupled with human-in-the-loop review, play a vital role in moderating online content to help create safe online spaces. Prediction models typically perform the initial review, with automatic decisions in cases of high-confidence, and less confident cases deferred to human review. As the adage goes, "garbage in equals garbage out" (Vidgen and Derczynski 2020): to train reliable data-driven machine learning models, we must first obtain high-quality training and testing data (Aroyo et al. 2022; Sambasivan et al. 2021).

However, annotation tasks like judging hate speech and classifying misinformation often require subjective judgement, which is potentially impacted by biases held by annotators (Aroyo et al. 2019). For example, Hube, Fetahu, and Gadiraju (2019) show how worker factors influence outcomes in a task that asks participants to label statements on controversial topics as 'neutral' or 'opinionated'. Similarly, a broader range of factors, such as skills (Kumai et al. 2018), cognitive ability (Hettiachchi et al. 2020), and worker environment (Gadiraju et al. 2017) influence task outcomes in both objective and subjective crowdsourcing tasks. In addition to worker factors, clarity of the task (Gadiraju, Yang, and Bozzon 2017), user interface (Alagarai Sampath, Rajeshuni, and Indurkhya 2014), and incentives (Singer and Mittal 2013) could also create biases. Whether originating due to worker factors, or inflicted by requester actions, these biases can significantly impact the annotation data quality with broader implications for models trained using labelled data (Bender et al. 2021; Wiegand, Ruppenhofer, and Kleinbauer 2019; Smith et al. 2022).

Significant prior work has examined agreement among crowd workers, comparing them with expert annotations, and exploring the relationships between worker characteristics and annotation quality (Poletto et al. 2021; Wiegand, Ruppenhofer, and Kleinbauer 2019). For example, recent work on toxic language annotation by Sap et al. (2022) investigates how annotator beliefs regarding race influence their toxicity ratings of African American English text. However, examining the relationship between questionnaire and annotation outcomes provides only limited insights into the worker's sense-making process, as well as how and whether they attempt to alienate personal beliefs to achieve consistency with task instructions and expectations.

To address this, we conducted two crowdsourcing studies on Amazon Mechanical Turk (MTurk) that combine quantitative and qualitative methods to better investigate how annotator biases may influence tagging of sexist and misogynistic hate speech on Twitter. While much prior work has explored worker biases in subjective judgements relating to hate speech tagging at large, our work specifically investigates misogynistic or sexist hate speech on the Twitter platform. Our first study gathers quantitative input from workers performing hate speech annotation. Results from 109 annotators show that annotator political inclination, moral in-

tegrity, personality traits, and sexist attitudes significantly impact annotation accuracy measured through agreement with expert annotations. Similarly, worker attributes have significant impacts on what portion of tweets they tag as hate speech. In our second study, we conducted semi-structured interviews with 9 crowd workers to garner further insights into how work practices and context, including employment conditions, can contribute to bias in subjective annotation.

Contributions. We make three key contributions. First, our study reveals a significant impact of political inclination, moral integrity, personality traits, and sexist attitudes of workers on their accuracy in annotating misogynistic hate speech. Second, through drawing on qualitative, semi-structured interviews with workers as they undertook their annotations, our findings offer new insights into the complex negotiations of the platform and personal conditions, structures and standards that annotation work requires. By highlighting how crowd work and the resulting annotations are influenced not only by worker factors but the structures under which they labour, we evidence the utility of a more holistic accounting of influential factors, and open the potential for more nuanced design considerations in the future. Finally, in collaboration with subject matter experts, we create and share a new Twitter dataset that contains misogynistic or sexist hate speech, including expert annotations of five subcategories of hate speech.

Related Work

Subjective Annotation Tasks

Annotation tasks span a wide extreme from *objective* tasks (*e.g.,* is there a dog in this picture?) to *subjective* tasks (*e.g.,* what is your favourite colour?), with a wide, varying-range of partially-subjective tasks (Nguyen et al. 2016) in between. Collecting annotations for subjective phenomena, in general, is challenging, and more so in crowdsourcing contexts involving temporary, contingent work between unknown parties with minimal communication channels.

Aroyo et al. (2019) discuss two types of subjectivity that can impact the worker judgement in toxicity annotation. First, subjectivity could be inherent to the topic or the domain, where workers could draw on personal preference or experience to make the judgement. In these tasks, two workers could simply have different opinions. Second, subjectivity could occur due to the ambiguity in input items or task instructions. Similarly, Sen et al. (2015) examine semantic relatedness judgements that ask workers to rate the strength of the relationship between two concepts. The study on MTurk reports that annotated datasets collected from 39 MTurk crowd annotators significantly vary from annotations collected from 72 scholars. Similar results were evident from work by Hube, Fetahu, and Gadiraju (2019), who found that crowd workers with strong opinions produce biased annotations, even evident among experienced workers. Furthermore, demographics, location, worker context, and work environment have been found to impact annotation quality in both subjective and objective tasks (Hettiachchi et al. 2021).

Mitigating Bias in Crowdsourced Annotation

Various approaches exist for detecting and mitigating biases stemming from crowdsourced annotations. While numerous post-annotation approaches exist to detect and/or mitigate algorithmic bias, it is preferable to avoid or reduce biases during the annotation process itself. Common data quality control measures, such as comparing worker labels to gold standard labels (*e.g.,* from experts), have typically been studied in the context of objective labelling tasks in which disagreement with gold can be construed as a labelling error. While the labelling error in objective tasks could represent the overall annotation quality, with subjective tasks, limited error measures cannot accurately cover the broader spectrum of subjective judgements involved in a task. Researchers have thus explored alternative methods, such as task design and presentation strategies. Hube, Fetahu, and Gadiraju (2019) experimented with a task where workers are given statements related to controversial topics and asked to tag them as 'neutral' or 'opinionated'. Their results show that social projection – or asking workers to label according to how they believe the majority of other workers would label them – leads to more consistent annotations. Similarly, "awareness reminders" ask workers to reflect on the controversial nature of the topics in the task and be mindful of the potential bias their personal opinions could have on their judgements.

Other approaches to mitigate bias require task requesters (*e.g.,* researchers or machine learning practitioners) to manually specify population requirements to mitigate bias. For example, Barbosa and Chen (2019) allow requesters to specify whether they need balanced or skewed populations with respect to specific worker characteristics. While such selections can help task requesters avoid certain worker groups depending on task needs, determining appropriate population parameters is not always straightforward. Task requesters themselves may also embody their own biases, complicating matters even further.

Implications of Precarious Work

Online crowd work, or data work more generally, has well-known challenges, leading to characterisation as *invisible work* (Irani and Silberman 2013) or *ghost work* (Gray and Suri 2019). While researchers and practitioners have sought a "gold mine" of inexpensive labelling of gold data, others have suggested that a "coal mine" metaphor may be more appropriate (Fort, Adda, and Cohen 2011). The negative impacts of these forms of labour on workers and their ethical consequences have a long history of such critique (Vallas and Schor 2020; Felstiner 2011). The most salient consequence of this, in regard to this particular study, is understanding the extent to which such precarious work can influence the resulting work products by undermining the quality and/or creating or amplifying biases in annotation. For example, the question of who is willing to undertake such work in general, or at the wages offered, is known to shape the demographics of the annotators who choose to participate in research studies or data labelling work.

The piecemeal nature of the work and its low pay rates ensure earning a living wage is challenging (Berg and Rani

2021; Naderi 2018; Alkhatib, Bernstein, and Levi 2017). Further, through breaking tasks down, and dispersing their completion, crowd work has enabled some firms to move away from reliance on in-house employment (Berg 2015), exacerbating the ethical considerations surrounding such modes of employment (Vallas and Schor 2020). These considerations are compounded by the power dynamics that frame this work, with task requesters able to reject and refuse payment for any tasks not completed to their satisfaction (Rea et al. 2020; Alkhatib, Bernstein, and Levi 2017). Qualitative research has been critical in highlighting these dynamics, particularly in drawing attention to how platforms such as MTurk "rely not only on the calculative mechanisms of control that metrics afford but also on normative mechanisms in the form of [...] inducements that strengthen user attachment" (Vallas and Schor 2020)(p. 279). Examining how these "hierarchical structures [...] inform the interpretation of data" (Miceli, Schuessler, and Yang 2020) (p. 2) is therefore critical.

Hate Speech and its Annotation

Hate speech annotation represents a family of subjective annotation tasks (Fortuna and Nunes 2018) essential to both offline data collection for model training and testing, as well as online, human-in-the-loop content moderation. Despite online platforms (*e.g.,* (Meta 2023; Twitter 2023)) and authorities having recognised hate speech as a growing concern, the definition of hate speech lacks global consensus, while many different definitions are used for related and equivalent concepts (Fortuna and Nunes 2018; Poletto et al. 2021). This lack of consensus, and the resulting variance in data annotation practices, also makes it more challenging to conduct rigorous benchmarking evaluations across publicly available datasets (Fortuna, Soler, and Wanner 2020).

As with other annotation tasks, bias in training data could lead to biased and inaccurate models (Bender et al. 2021; Wiegand, Ruppenhofer, and Kleinbauer 2019; Smith et al. 2022). Wiegand, Ruppenhofer, and Kleinbauer (2019) highlight that under more realistic settings, abusive speech classification performance is much lower than what is reported in previous research. Similarly, recent work shows that agreement among annotators strongly correlates with hate speech recognition quality in automated methods (Kocoń et al. 2021). Rahman et al. (2021) report a significant drop in hate speech detection accuracy of current models when tested on a new hate speech dataset that includes broader forms of hate speech. Sap et al. (2022) examine how the annotator beliefs and identities inflict biases in toxic language detection tasks. In particular, Sap et al. report notable variations in judgements considering conservative annotators and ones scoring higher values for racist beliefs. In their task, such annotators were less likely to rate anti-Black language as toxic, but more likely to rate African American English as toxic. Salminen et al. (2018b) report on a study where crowd workers of 50 countries provided toxicity annotations for social media comments. The reported hate interpretation (*i.e.,* rated intensity of hatefulness) scores differ significantly between the countries considered, but the interpretation differs more by individuals than by countries. Salminen et al. (2018b) note the importance of considering user attributes in the hate speech annotation processes and studying their impact.

A wide variety of hate speech datasets now exist (Poletto et al. 2021; Vidgen and Derczynski 2020; Davidson et al. 2017). Poletto et al. (2021) review 64 datasets (56 annotated corpora and 8 lexica), with only 9 including sexist or misogynistic content. In terms of sources, 32 datasets contain tweets, while others include comments from social platforms like Facebook, Reddit, and YouTube, as well as news websites. The annotation process involves experts (*i.e.,* judges with subject knowledge), or non-expert volunteers or annotators recruited through crowdsourcing platforms. In general, broad coverage hate speech datasets typically contain a relatively small amount of misogynistic or sexist hate speech.

Guest et al. (2021) propose a taxonomy for classifying misogynistic hate speech and a dataset based on Reddit posts and comments. The paper defines four categories for misogynistic content: *Misogynistic Pejoratives*, *Descriptions of Misogynistic Treatment*, *Acts of Misogynistic Derogation*, and *Gendered Personal Attacks Against Women*. Categories for non-misogynistic content include *Counter-Speech Against Misogyny*, *Non-misogynistic Personal Attacks*, and other content that doesn't belong to any of the categories. While many other hate speech taxonomies have been proposed in the literature (Fortuna and Nunes 2018), they provide limited coverage on specific sub-categories needed to distinguish different types of misogynistic hate speech (Guest et al. 2021).

Increasing attention is being devoted to how annotator attributes influence the annotation decisions in assessing hate speech at large (Al Kuwatly, Wich, and Groh 2020; Keswani, Lease, and Kenthapadi 2021; Kumar et al. 2021; Sap et al. 2022; Gordon et al. 2022). However, relatively little work has explored the impact of annotator attributes on sexist or misogynistic hate speech annotation decisions in particular (Wojatzki et al. 2018; Goyal et al. 2022). Wojatzki et al. (2018) report a study with 40 men and 40 women judging 400 German assertions about women, and shows that both men and women judge hate speech consistently, particularly for cases of extreme misogyny. Civil Comments (Borkan et al. 2019) provides toxicity labels for 1.8M news comments, with 450K comments also labeled for the demographic group targeted (*e.g.,* gender). In forthcoming work, Goyal et al. (2022) augment the Civil Comments dataset with new annotations for annotator demographics, thereby allowing the study of how annotator demographics impact labelling decisions for different target groups.

Study

To better understand how annotator biases may influence annotation of sexist and misogynistic hate speech on Twitter, we conduct two crowdsourcing studies on MTurk that combine quantitative and qualitative methods. First, we gather sexist and misogynistic hate speech labels from crowd annotators in order to measure effects quantitatively. Following this, we conduct semi-structured interviews with nine crowd workers to garner further insights into their work context and thought processes. We first discuss common elements across

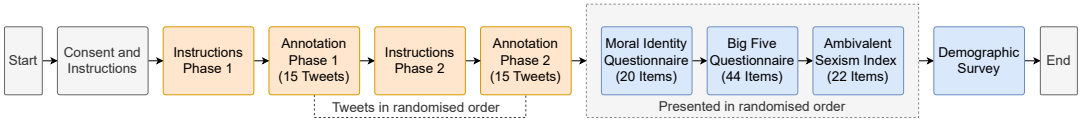

Figure 1: Flow of Study 1.

the two studies, then present the details of each study in the following sections. Both studies were approved by our human subjects oversight board.

Sampling Tweets for Annotation

Hate speech is a broad term that refers to content targeting a person or group of people on the basis of their protected characteristics (*e.g.,* race, ethnicity, national origin, disability, religious affiliation, sexual orientation, sex, gender identity). In this work, we consider annotation of sexist/misogynistic hate speech or hate speech directed at women on Twitter, a specific target group which has received far less attention than the study of hate speech annotation at large.

A subject expert on anti-discrimination, and free speech and media law curated a list of 16 Twitter profiles (public figures) susceptible to sexist and misogynistic hate speech. We also contacted these individuals to obtain their consent to include their Twitter handle and relevant tweets in our research. We received positive responses from seven individuals, and collected the 5,000 most recent tweets for each profile. The selected individuals included authors, journalists, and activists from White, African-American, and Asian racial groups located in the US, UK, and Australia. Their profiles had 194,308 followers and 26,098 tweets on average. To avoid ambiguity on whether a tweet is directed at the specific person, we excluded any tweet that included more than one mentions (*i.e.,* twitter handles in the tweet). Finally, we ordered by the tweet date and selected 2000 recent tweets for each profile.

Annotation Guidelines and Expert Labels

Prior work has examined the utility of categories/factors, rating scales and comparative judgements when collecting hate speech annotations (Poletto et al. 2019). While rating scales are reported to be easy to use for annotators, they produce results which are difficult to understand and lead to poor overall performance in automated models. Comparative judgements such as best-worst-scale lead to better model outcomes. However, such ranking input scales are not suitable for our study design. Therefore, we collect categorical responses in our study. As we are interested in examining a specific type of hate speech, we use a tailored hate speech taxonomy instead of using generic and broader hate speech taxonomies (Salminen et al. 2018a) that have limited and incomplete label categories for misogynistic hate speech. Specifically, De Silva (2020) presents a functional theory of sex-based vilification, and Guest et al. (2021) propose a taxonomy, from which we adapt the following categories of hate speech:

- *Threats and Violent Abuse (TVA)* are particularly serious examples of sex-based hate speech. It includes death and rape threats, as well as violent 'correctives' which tell the participant that they 'deserve/need [to be violated]'
- *Sexualised Abuse (SA)* reduces women to tools or objects for men's sexual use or pleasure. It exposes women's sexuality in public, irrelevant, humiliating, and/or distressing ways.
- *Other Objectifying Speech (OOS)* is speech that is less severe than threats and violent abuse or sexualised abuse, but that still treats women as for use by others (typically by men). It is speech that treats women as valuable only if they are attractive or useful to others.
- *Other Contemptuous Speech (OCS)* communicates hatred, dislike, or disrespect for women that does not fit the categories described above. Other contemptuous speech typically treats women as inferior to men.
- *Other Hate Speech (OHS)* is content that you think ought to be tagged as hate speech, but that does not fit into any of the above categories of sex-based hate speech.

Two authors (one woman, and one man) independently annotated the selected collection containing 14,000 tweets according to the annotator instructions and examples generated for the annotation Phase 2 (*i.e.,* five hate speech categories and not hate speech). Afterwards, a third author, the expert mentioned earlier, reviewed the annotations, discussed any conflicts with the other two authors, and determined the final gold annotations. For the final dataset, we selected all the tweets tagged as hateful, and randomly selected a subset of tweets tagged as not hateful. Our final dataset includes 140 tweets, with 90 tagged as not hateful and 50 tweets tagged as hate speech (Threats and violent abuse - 4, Sexualised abuse - 9, Other objectifying speech - 7, Other contemptuous speech - 10, Other Hate Speech - 20). While we treat these expert annotations as gold data to evaluate the agreement between annotators and experts, we acknowledge that expert annotations are not equivalent to ground truth due nuanced nature of defining hate speech (Fortuna, Soler, and Wanner 2020).

Study 1: Collecting Hate Speech Annotations

The first MTurk study compared hate speech labelling decisions of crowd workers under two variant designs: 1) binary labelling ('Hate Speech' or 'Not Hate Speech') vs. 2) fine-grained labelling between the five specific types of hate speech described in Section vs. 'Not Hate Speech'. As illustrated in **Figure 1**, each worker first performed binary labelling of 15 tweets (Phase 1), and then performed fine-grained labelling of another, different set of 15 tweets (Phase 2). For each available category, task instructions both described the category and provided two example tweets. Workers were shown the instructions before beginning each

phase. Instructions could be reviewed at any time during annotation by clicking the 'Show Instructions' button.

Following annotation, workers were presented with three questionnaires (in randomised order) and a final demographic survey, as shown in Figure 1. Similar to Hube, Fetahu, and Gadiraju (2019)'s "awareness reminders", some prior work has shown that asking participants about their demographics at the start of a study can increase their self-awareness and influence their later responses, which is why we left this demographic survey for the very end. Self-reported demographic attributes collected include gender, political identity, the average weekly income, and crowdsourcing income as a percentage of primary income.

The study included three questionnaires. The 44-item *Big Five Personality* questionnaire (Goldberg 1992) has been used in prior research that investigates annotation quality in crowd work (Lykourentzou et al. 2016; Kazai, Kamps, and Milic-Frayling 2012). It measures five personality traits: Extroversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. More pertinent to our specific hate speech annotation task, we also included the *Moral Identity* questionnaire (Black and Reynolds 2016) and the *Ambivalent Sexism Inventory* questionnaire (Glick and Fiske 1996). The Moral Identity questionnaire primarily captures how people make moral choices and has two dimensions: moral identity and moral self. Moral identity measures "the desire to make intention and action consistent, and how much value participants place on acting according to moral principles.". Moral self measures "how closely participants identify with moral values". The Ambivalent Sexism Inventory measures hostile sexism and benevolent sexism (Glick and Fiske 1996).

The task was deployed for MTurk using psiTurk (https://psiturk.org/). As pre-qualifications, workers were required to have completed more than 1000 tasks, have an approval rate greater than 95%, be located in the US, and have MTurk's adult qualification. For quality control purposes, we also included two attention-check questions in the annotation task. This led to the exclusion of responses from 9 workers who failed at least one of the checks.

Study 2: Conducting Semi-structured Interviews

Our second MTurk study draws on previous approaches to ethnographic research with crowd workers such as that outlined by Gray et al. (2016) and Williams et al. (2019). We conducted semi-structured interviews via remote video meetings. Similar to Study 1, workers were required to have completed more than 1,000 tasks, have an approval rate greater than 95%, be located in the US, and have MTurk's adult qualification. In addition, we did not allow any workers who had completed Study 1 to also attempt Study 2. Workers initially expressed interest and reserved a time slot through a qualification task and then started the interview task during the agreed time slot.

Following Gray et al. (2016), we asked workers to "demonstrate how they [do] their crowd work" by sharing their screen with the researcher and talking through the decisions that underpin their work practices. They were asked to describe how and why they make certain decisions to better understand their thought processes, whether individually or structurally determined. Interviews lasted approximately 45 minutes, and participants were compensated $15 USD.

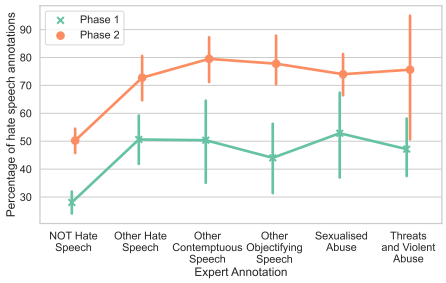

Figure 2: Percentage of hate speech annotations received for tweets categorised by the expert annotation.

Results

Quantitative Results

We use two standard measures in our analysis. First, we assess whether crowd worker annotations agree with our expert annotations, termed as "accuracy" when considering expert annotations as gold standard data. Second, we examine the "percentage or the number of hate speech annotations" provided by each annotator or received by a collection of tweets. We collected 8-12 annotations per tweet, with a total of 3,270 responses collected from 109 workers. On average, each worker spent 18.4 ($SD = 10.4$) minutes on completing the full task. Workers spent 14.3 ($SD = 5.2$) seconds per tweet on average in Phase 1. In Phase 2, the average annotation time was 15.2 ($SD = 2.3$) seconds.

Identifying Hate Speech How well do workers identify whether a tweet should be tagged as hate speech? For each tweet, we calculated the percentage of workers who rated it as hate speech (considering only binary hate / non-hate distinctions). **Figure 2** shows the percentage of hate speech annotations received for each tweet group, categorised according to the expert annotation.

In addition, to measure inter-annotator agreement, considering the nature of our crowdsourced data collection task, we use Krippendorf's alpha (Hayes and Krippendorff 2007). Unlike other inter-annotator agreement measures, Krippendorf's alpha is suitable when each annotator only labels a subset of total items (Hayes and Krippendorff 2007). When considering binary annotation outcomes, our results indicate a Krippendorf's alpha score of 0.149 in Phase 1 and 0.161 in Phase 2. While these scores only indicate slight agreement, it is a common observation in hate speech annotation tasks (Fortuna and Nunes 2018).

Categorising Hate Speech In the second phase of annotation, workers were expected to label the specific type of hate speech. Workers found it challenging to do so. As detailed in **Table 1**, for example, for tweets with an expert annotation of 'Threats and Violent Abuse (TVA)', only 9 (23.7%) out of 38 annotations accurately labelled the specific category. Similarly, for 'Other Contemptuous Speech (OCS)', only 33 (30.8%) out of 107 annotations correctly classified

Crowd Worker Annotation		Expert Annotation					
		NHS	OHS	OCS	OOS	SA	TVA
Phase 1	NHS	702	105	54	46	47	25
	HS	268	112	55	36	54	22
Phase 2	NHS	496	62	22	16	26	8
	OHS	128	52	17	11	9	7
	OCS	151	42	33	14	16	5
	OOS	103	34	17	16	26	4
	SA	63	7	11	8	16	5
	TVA	50	21	7	6	8	9

Table 1: Total number of annotations received for tweets grouped by the expert annotation.

the tweets. When considering all the categories in Phase 2, the resulting Krippendorf's alpha score of 0.086 also indicates very low agreement among workers.

Relationship with Questionnaire Outcomes We calculated annotation outcomes for each worker to examine the impact of worker factors. Shapiro-Wilk tests indicated that accuracy and hate speech annotation percentage scores do not follow a normal distribution. Spearman rank correlation coefficients between questionnaire outcome scores, and worker accuracy and percentage of hate speech annotations provided by each worker are reported in **Table 2**. We observe strong negative correlations between moral integrity and worker accuracy in Phase 2. For moral self, the results indicate a moderate positive correlation with Phase 2 accuracy. When examining relationships with the big five dimensions, three attributes (Agreeableness, Conscientiousness and Openness) have moderate positive correlations with Phase 2 accuracy. Furthermore, a moderate negative correlation is evident between benevolent sexism and worker accuracy in Phases 1 and 2.

When considering the percentage of hate speech annotations, in most cases, the correlation scores are in the opposite direction compared to accuracy. Between moral integrity and the percentage, strong positive correlations are evident in Phase 2, while a moderate positive correlation exists in Phase 1. We note a moderate negative correlation between moral self and the percentage in Phase 2. With the big five dimensions, Agreeableness and Conscientiousness have moderate negative correlations with Phase 2 accuracy. Moderate positive correlations are also evident between benevolent sexism and the percentage in both phases.

Impact of Demographic Factors Our results indicate a significant impact of worker political leaning on the hate speech annotation accuracy. As shown in **Figure 3a**, the difference is evident in annotations collected in both Phase 1 and Phase 2. In Phase 1, a Kruskal-Wallis test indicates that the worker accuracy differed over political leaning, $H(2) = 14.2, p < 0.001$. A post-hoc Dunn test with Bonferroni adjustment shows that worker accuracy for left-leaning workers ($N = 39$, $M = 70.3$, $SD = 11.5$) is significantly higher compared to both centre ($N = 23$, $M = 57.5$, $SD = 10.7$, $p < 0.001$) and right ($N = 47$, $M = 62.6$, $SD = 15.6$, $p < 0.001$) leaning workers. There is no significant difference among the centre and right-leaning annotators. Similar results are evident in Phase 2, with the Kruskal-Wallis test showing a significant difference in accuracy based on the political leaning, $H(2) = 19.6, p < 0.001$. A post-hoc Dunn test with Bonferroni adjustment indicates a significantly higher accuracy for left-leaning annotators ($M = 68.7$, $SD = 15.7$) compared to centre ($M = 52.5$, $SD = 15.6$, $p < 0.001$) and right ($M = 53.6$, $SD = 17.7$, $p < 0.001$) leaning workers, while the difference among the centre and right workers is not significant.

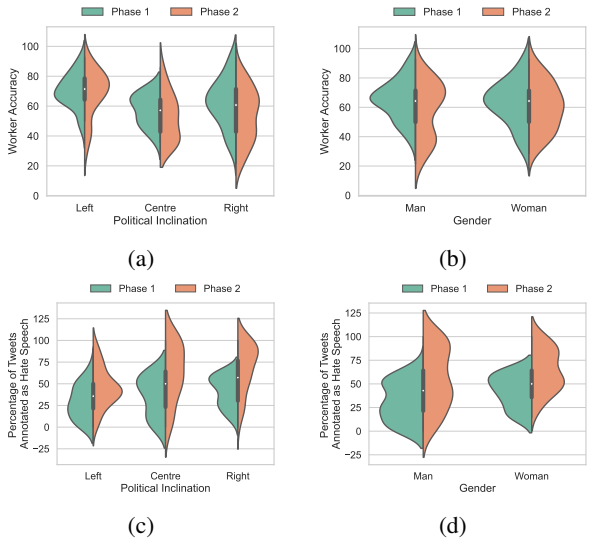

Figure 3: Impact of worker political inclination and gender on annotation outcomes.

In Phases 1 and 2, worker political leaning also significantly impacts the percentage of hate speech annotations provided (**Figure 3c**). In Phase 1, a Kruskal-Wallis test indicates that the percentage of hate speech annotations significantly varied over worker political leaning, $H(2) = 8.4$, $p < 0.05$. A post-hoc Dunn test with Bonferroni adjustment shows that the percentage of hate speech annotations provided by left-leaning workers ($M = 28.9$, $SD = 19.7$) is significantly lower than right-leaning workers ($M = 41.5$, $SD = 17.8$, $p < 0.05$), and the differences among centre-right and centre-left are not significant. In Phase 2, a Kruskal-Wallis test indicates significant differences ($H(2) = 15.5, p < 0.001$), and a post-hoc Dunn test with Bonferroni adjustment shows that left-leaning workers ($M = 45.4$, $SD = 22.5$) marked a lower percentage of tweets as hate speech compared to both centre ($M = 62.1$, $SD = 32.6$) and right ($M = 68.1$, $SD = 27.7$) leaning workers. The difference between centre and right workers is not significant.

As seen in **Figure 3b**, we further examine whether the self-reported gender of the worker impacts hate speech annotation accuracy. Mann Whitney U tests indicate no significant differences in accuracy based on the gender of the worker in both phases. In Phase 1, men ($N = 65$, $M = 65.1$, $SD = 13.6$) demonstrated a slightly higher mean accu-

Test	Dimension	with Accuracy			with % of Hate Annotations		
		Phase 1	Phase 2	Overall	Phase 1	Phase 2	Overall
Moral Identity Questionnaire	Moral-Self	0.234	**0.384**	**0.384**	-0.154	**-0.360**	**-0.312**
	Moral-Integrity	**-0.480**	**-0.551**	**-0.607**	**0.368**	**0.614**	**0.567**
Big Five Personality	Extraversion	**-0.327**	-0.152	-0.271	0.286	0.183	0.226
	Agreeableness	**0.302**	**0.443**	**0.459**	-0.257	**-0.470**	**-0.435**
	Conscientiousness	0.175	**0.381**	**0.350**	-0.220	**-0.484**	**-0.428**
	Neuroticism	-0.083	-0.154	-0.129	0.123	0.283	0.250
	Openness	0.111	**0.338**	0.298	-0.158	**-0.326**	-0.286
Ambivalent Sexism Inventory	Hostile	-0.271	-0.057	-0.173	0.117	0.125	0.132
	Benevolent	**-0.446**	**-0.364**	**-0.467**	**0.337**	**0.496**	**0.477**

Table 2: Spearman rank correlation scores between questionnaire outcomes and hate speech annotations.

racy compared to women ($N = 44$, $M = 63.1$, $SD = 15.2$), whereas in Phase 2, mean annotation accuracy for men ($M = 57.9$, $SD = 19.2$) was lower than for women ($M = 60.0$, $SD = 16.3$). In Phase 1, a Mann-Whitney U test indicates a significantly higher percentage of tweets tagged as hate speech by women ($N = 44$, $M = 42.4$, $SD = 17.1$) compared to men ($N = 65$, $M = 31.4$, $SD = 21.0$) ($U = 995.5$, $p < 0.01$). The difference is not statistically significant in Phase 2 (**Figure 3d**).

Qualitative Results

In this section, we draw on qualitative data collected through semi-structured interviews with nine MTurk workers. Participant demographics are shown in **Table 3**. In exploring the sensemaking processes of online crowd workers, we were especially inspired by Miceli, Schuessler, and Yang (2020)'s investigation of the work of data annotation for computer vision, structured around three questions: "How do data annotators make sense of data? What conditions, structures and standards shape that sense-making praxis? Who, and at what stages of the annotation process, decides which classifications best define each data point?" (p. 2). The qualitative component of our research set out to address similar guiding questions: How do crowd workers make sense of hate speech? What conditions, structures and standards shape that sense-making? and How, and at what stages of the annotation process, do crowd workers decide what comprises hate speech?

Interviews took approximately 45 minutes and were conducted via remote video meetings. The first half of the interview covered questions regarding each worker's previous experience of crowd work in general, and annotation tasks in specific. In the second half, the worker was invited to 'demonstrate how they [do] their work' (Gray et al. 2016) by sharing their screen and talking through the decisions that underpin their work practices. Participants were asked to describe how and why they make certain decisions to better understand the context underscoring their decision-making, and whether and how these contexts were individually or structurally determined.

ID	Gender	Sex	Age (years)	Crowd worker for
P-1	Man	Male	18-30	More than 4 years
P-2	Man	Male	30-40	More than 4 years
P-3	Man	Male	30-40	1-2 years
P-4	Woman	Female	40-50	Less than 1 year
P-5	Woman	Female	30-40	2-3 years
P-6	Man	Male	60-70	More than 4 years
P-7	Man	Male	18-30	Less than 1 year
P-8	Woman	Female	40-50	More than 4 years
P-9	Man	Male	18-30	1-2 years

Table 3: Demographics of the participants.

With participant consent, interviews were recorded, including both audio and visual feeds. Participants were invited to switch their camera off if they desired, but the interviewer kept their camera on throughout the interview. Following Braun and Clarke (2006), we undertook a thematic analysis, identifying descriptive themes or patterns within the qualitative data. Our analysis was inductive (driven by the data) rather than deductive (theory driven). As such, we used an open coding approach (Holton 2007).

Our analysis was multi-staged. First, one of the authors familiarised themselves with the data by reading and rereading the transcripts. They then coded the data, looking for themes and patterns. Third, they collaborated with another author to review and refine the codes into larger, descriptive themes. In undertaking this process collaboratively, we sought not to seek replicability or objective accuracy, but rather internal consistency and representation of experiences within and across the corpus.

In what follows, we show how workers made sense of the task, and how this sensemaking was shaped by complex negotiations between subjective motivations, platform structures, task instructions, and external contextual factors. In doing so, we highlight the ways in which the outcomes of crowd working are influenced by the structures under which they labour. In making this argument, we move away from the existing focus on subjective influences of worker characteristics and their impact on biases stemming from annotator beliefs, characteristics, and demographics, and towards a more holistic accounting of influential factors.

Defining hate speech In discussing how they defined hate speech, and thus categorised the tweets, participants frequently opined on the importance of free speech, drawing on geographically and culturally bound understandings of hate speech. According to P-3 (*i.e.,* Participant 3), for instance, "*I think you should be able to say anything you want for the*

most part, as long as it's not threatening [... I'm] a pretty free speech type of person, to be honest". Likewise, P-4 said, "*I'm more in favour of free speech. So even if it might be offensive to people, I think people should still be able to say it*". As P-4 continued: "*I feel like people can say stuff, in America at least*". P-4 stated "*[...] In the USA, we have different laws, I believe, for hate speeches*". In these instances, definitions of hate speech were geographically (American) and culturally (Western) bound. For P-9, these influences were evidence of the collective generation of such definitions: "*we're all influenced by each other, whether we like it or not*". The importance placed on free speech directly influenced how participants approached annotation tasks. As P-8 explained: "*I would prefer to train it [through annotations so] there's like leeway, not everything is meant to be offensive, even if it accidentally is*". P-9 echoed this sentiment, noting that "*hate speech is too ambiguous. It's too all-encompassing and it's like a thinly veiled attempt at censorship, and it's clearly taking sides*". To deviate from this view, and to classify something as hate speech, then, P-3 told us that "*you have to [be able] to say there's definitely, the person hates this person*", or that there was clear evidence of "*hate speech towards, like, a race or a gender*", or what P-8 described as "*a protected class type vibe*".

Other participants reiterated similar views and told us how previous experience with similar annotation tasks, and/or awareness of legal approaches to understanding hate speech informed their work. P-4, for example, told us that: "*I've kind of been conditioned by the past tasks to assume that true hate speech should have implicit or explicit calls to violence, or be a little bit more than just politically incorrect*".

Fluctuations in hate speech sensemaking occurred when such 'standard' definitions of hate speech encountered personal values or subjective ethos. P-1, for example, used empathy to interpret whether a tweet was hate speech or not: "*I'm of the belief that we don't need to have something happen to us to empathise with whether or not it's right or wrong*". These values influenced annotating practices to varying degrees. For P-8, personal values took precedence over following task instructions or established definitions, sharing that "*Sometimes I get in trouble for it, but I think it's worth it. [...] I mean there's stuff worth planting your feet in the ground for*". For some crowd workers, their subjectivity was an inherent part of their annotation work, with this something they presumed task requesters were taking into account. As P-6 told us, "*I bring my experience to the table, and I figure that's part of what you're paying for*". For this participant, tasks had to be congruent with their personal values, or else they would "*sign out of them and go on something else*" altogether.

At multiple stages throughout the experiment, participants encountered 'borderline' cases that introduced significant ambiguity. In these instances, a range of different values and information-based strategies were employed. Those guided by values were most likely to err on the side of caution to mitigate potential instances of hate speech. As P-6 stated, "*My feeling about hate speech is if it makes somebody uncomfortable, why do it? And if you're doing it after somebody is uncomfortable, then it's on you. And so you err on the side of caution.*" In this instance, the participant's values oriented their consideration of others, and led to their tendency to lean towards categorising tweets as hate speech, rather than not.

Weighing up financial imperatives, task instructions, and subjective values The influence of financial imperatives on the sense-making practices of our participants cannot be overstated. For all nine of our participants, the use of MTurk was principally driven by financial incentives. P-3, for instance, had taken up crowd working when one of their existing three sources declined: "*when I first started it MTurk I had just wanted another source of income, because I had three sources, [and] one became less. So, I'm like, okay, I'll check it out [...] it's a nice little source of a little extra [income], not the greatest, but it's okay*".

Despite the strong influence of personal values for some participants just described, in most instances, task instructions were prioritised. For example, while P-4 noted that while their "*subjective reasoning*" underlay the sense-making that informed their annotations, these were superseded by the task instructions: "*the task instructions are always my primary goal*". The primacy of task instructions was driven by financial imperatives: "*At the end of the day, this is a job I'm getting paid in actual physical currency. So, I do try to do a good job as much as possible*". As P-4 explained: "*they [task requesters] are not looking for my opinion. They have a set of guidelines, and they want their data in those buckets. And I do take this job seriously [...] I don't want to screw up someone's data set*". If participants deviated from the task instructions, they told us, they risked not receiving payment for their labour. This was reiterated more strongly by others, like P-9, who told us that "*being subjective is a big no-no for me on MTurk because there's a lot of people who don't pay you for your batches due to subjective criteria*".

For some participants, task instructions were followed even if these were at odds with personal and subjective values: "*usually, they have rules on what they consider hate speech in the instructions. [...] So you follow their rules. Do I think it's always hate speech? No, but uh, you follow the rules*" (P-8). P-9 adopted a similar outlook in stating that in doing annotation tasks, "*I'm gonna do what my overlord's telling me to, and I'm gonna get my money and I'm gonna be on my way*". A notable exception was P-6, who articulated that due to their experience and accumulated status on the platform, they could adopt more of a subjective viewpoint, "*You've got a bit of power in the system to have your own stance, I suppose*".

The details included in task instructions were, therefore, critical for participants to negotiate tensions between their subjective values and their need to get paid. For example, for P-1, if the task instructions said simply: "'*annotate hate speech', then I'm gonna be like, 'okay, what do I think of hate speech as?' But if it says, 'annotate hate speech, here's your definition, here is the lens you should be looking at it through', then I would obviously follow the instructions there.*". Broader understandings of the type of annotations expected by requesters also influenced how workers inter-

preted instructions, as identified by P-2, "*If I'm unsure, I revisit the criteria that's been laid out*". For P-8, these kinds of tasks required additional focus. As they explained, when working on annotation tasks with specific guidelines, they would "*end up getting out my handy dandy notebook and making notes of what the instructions are*". This was because, "*if you don't follow instructions, you won't get any more work from that requester, or your task could be rejected*" (P-8).

Reflecting on the lack of definitive instructions in Phase 1, P-4 commented that they would usually "*feel uncomfortable*" at this point and would "*message the [task] requester and ask them for more guidelines*". According to this participant, achieving this clarity was critical: "*the more communication, the more successful the project will be*". This was due to prior experiences, where task requesters had provided insufficient guidance, and the resulting annotations were rejected. For P-4, clarity around task instructions was critical to ensuring requesters received "*high-quality data*", and that the crowd worker received payment for their labour. As articulated by P-8, "*I mean, it's whatever you want, you're paying for it, so you should get what you're paying for.*"

Filling in the gaps All nine participants commented on the difficulty of the task at hand, specifically referring to the challenge of categorising something as hate speech with minimal context, such as the surrounding conversation, or information regarding the sender and recipient. This was especially so in instances of 'borderline' cases. As P-3 explained, "*I don't know the context, so it's a little hard*". For P-8, decision-making in these instances was further challenged by what they described as the combination of their age and little social media presence. As a result, they were sometimes unsure if a word or phrase was offensive or not. To negotiate this, they described workaround practices, such as visiting websites like Urban Dictionary, where they would endeavour to decipher what was being said. As they explained, "*I'm not on social media a whole lot, I felt like, maybe these are terms that people are using [. . .] maybe I'm out of the loop*". Participants also noted that their ability to effectively complete such tasks relied on their keeping up with evolving understandings of hate speech. As reflected by P-2, "*like ten years ago, that might not be classified as hate speech, but now that might be [. . .] things have changed*".

The gap between the subjectivity of the individual crowd worker and the recipient of the hate speech also requires negotiation through the annotation process. While P-3 noted that although the text of the tweet in question "*doesn't really irritate me [. . .] it feels like hate speech just because it feels really aggressive*". In this instance, P-3 told us, they would have preferred to categorise this instance within a third category to indicate its borderline status. Given, however, that this stage of the study allowed participants only to categorise tweets as 'hate speech' or 'not hate speech', P-3 ultimately decided "*I'd say hate speech*". In later stages of the study, the examples provided within the task instructions, were seen as useful for demystifying ambiguous cases. As P-1 explained, "*a set of examples is always helpful*". Moving beyond the influence of task instructions and the subjective experience of the crowd worker, P-8 also took into account societal discrimination they "*observe in the world*" in determining whether something was hate speech or not.

To aid with this process of discernment, multiple participants described looking for key words or a "*handful of phrases or a handful of keywords that jump out at you*" (P-1). In a similar vein, P-4 told us that they "*look for specific words*". Re-reading for context surrounding the perceived hate speech served a similar purpose for P-2: "*[. . .] this section right here jumped out for me right away, and I went and reread it again to understand the full context*". Referring to task instructions also provided participants with scaffolding for understanding and categorising tweets as hate speech or otherwise. P-4, for instance, revisited the task instructions to remind themselves of what "*specific racial slurs*" were referred to. While the task instructions were useful for this participant in terms of filling in the gaps, for others like P-4, the additional detail introduced further challenges and increased the time required to complete the task: "*obviously, there's going to be more thinking that goes to it*".

Accounting for contextual factors Beyond the influence of subjectivities and platformed structures such as financial incentives, participants also told us how their annotations could be influenced by external contextual factors, like the mood that they were in that day. As P-3 explained, "*Depending on my mood that day, I might be 'I hate how sensitive the whole world is today', and I might be like, 'okay, I'm [going to select] non-hate speech'. It could be my mood, something I read, [or] just how I feel about things that day that could affect the way I'm answering*".

Annotation tasks could have affective impacts on the crowd workers themselves. P-4, for instance, told us about hate speech annotation tasks that they had worked on previously. Unlike in other tasks, where they "*would just keep going*", with hate speech annotation, they had to "*walk away from my computer every 30 minutes or so. Because even if you can be emotionally disconnected, it just starts to fatigue you because it's exhausting*". Such emotional labor underlying both volunteer and commercial moderation work has been well-documented (Dosono and Semaan 2019; Gilbert 2020; Roberts 2019; Steiger et al. 2021; Wohn 2019).

Discussion

Annotations and Worker Characteristics Our results indicate a significant relationship between crowd worker annotation accuracy and worker characteristics. As detailed in Table 2, we note a strong negative correlation between worker moral integrity, measured by the moral identity questionnaire and hate speech annotation accuracy. Moral identity indicates one's desire to make moral intentions and actions consistent, in other words, equality of private and public action. Guided by our qualitative findings in the sub-section on financial imperatives, we argue that following task instructions and providing accurate (*i.e.,* expert-aligned) subjective judgements in hate speech annotation tasks require workers to suppress their values, leading to the observed negative correlation. The impact of moral identity also aligns with work by Hube, Fetahu, and Gadiraju (2019)

where workers with strong opinions provided biased labels.

In prior work, the personality of crowd workers is known to be associated with objective task performance (Lykourentzou et al. 2016). Similarly, our results indicate moderate positive correlations between Phase 2 accuracy and agreeableness, conscientiousness and openness. In line with our findings, Sang and Stanton (2022) reported that annotators' big five personality factors, including agreeableness and conscientiousness, influence how workers judge hate speech annotations of 'showing fear', which include misogyny, homophobia, and xenophobia. Our results related to Ambivalent Sexism Inventory suggest that a higher score of benevolent sexism (*i.e.,* valuing stereotypical attributes in women and believing that women need to be protected) could lead to tagging more content as misogynistic hate speech and subsequently reducing the overall task accuracy.

Furthermore, our findings related to annotator gender agree with previous work finding no significant differences when men and women annotated hate speech directed at women (Wojatzki et al. 2018; Goyal et al. 2022). However, women tend to mark more content as misogynistic hate speech than men in Phase 1, where additional categories, specific guidelines and examples were unavailable.

Our findings are valuable to pre-select workers or assign tasks (Hettiachchi et al. 2020; Barbosa and Chen 2019) to obtain expert-aligned outcomes when crowdsourcing hate speech annotations. However, we note that pre-selection based on demographics without evidence could lead to further biases. Thus, we recommend informed pre-selection and recruiting diverse crowds considering other factors.

Subjective Judgements Analysis of qualitative data from semi-structured interviews with crowd workers provides useful nuance to further understand the quantitative data just discussed. Annotation outcomes, and the process taken to achieve them are not only impacted by worker factors but simultaneously shaped by the structures under which they labour. In this way, our work further strengthens findings around crowd work that evidence the impact of structures and hierarchy on both crowd work labour and outcomes (Miceli, Schuessler, and Yang 2020).

The structures and standards, such as financial incentives and task instructions under which crowd workers undertake content moderation tasks shaped the process of defining hate speech, and the processes workers took to complete different tasks. For example, platform approval ratings, projected task earnings, fear of tasks getting rejected and understandings of what specific task requesters were looking for prompted workers to prioritise task instructions over personal values. While subjective values, empathy and altruistic motivations influenced some participants' labour (Miceli, Schuessler, and Yang 2020), the financial incentives for successfully aligning annotation outcomes with task requester expectations superseded individual values for the majority.

Similar to objective tasks (Gadiraju, Yang, and Bozzon 2017), clarity in task instructions and examples appear to play a vital role in subjective judgements. Our quantitative results in Figure 2 revealed that the percentage of tweets accurately tagged as hate speech in Phase 2 is higher compared to Phase 1. In our qualitative findings, we highlighted that workers expect detailed task instructions for subjective tasks. However, workers are also primarily driven by financial needs and want to maximise the rewards they receive for work duration. Therefore, task requesters need to be mindful of the additional time required to read and understand instructions. While more detailed instructions can cover more borderline cases, there is a potential risk of workers ignoring or skimming them if they are too time-consuming.

Surfacing the complexity of these negotiations helps to clarify the structural influences that impact the crowd working process and resulting hate speech annotations. In many of the cases just described, these structures overshadowed the impact of individual worker characteristics. Moreover, hate speech sensemaking was culturally and geographically bound, with legal definitions and broader personal expectations affecting workers' interpretations of hate speech (Miceli, Schuessler, and Yang 2020). Although crowd work is often discussed in terms of its geographic dislocation (Elbanna and Idowu 2022), our findings indicate that geography may remain relevant in terms of defining hate speech. In outlining these findings, we indicate the potential for a more holistic accounting of influential factors on crowd work annotations, and raise critical questions concerning the links between task requesters, platform structures, and worker practices, that underpin crowd work outcomes.

Limitations We note three key limitations in our work. First, avoiding potential biases and increasing the variety of hate speech in the dataset was an important consideration (Rahman et al. 2021). Thus, we avoided an approach like keyword search and used a set of seed profiles. However, with the current process, our hate speech dataset only consists of tweets relating to seven public figures. Second, we recruited only nine participants for the qualitative study due to challenges in recruiting crowd workers for interview tasks. Third, in a crowdsourcing setup, it is challenging to validate self-reported demographics and questionnaire outcomes (Ipeirotis 2010; Marshall and Shipman 2013; Ross et al. 2010). While we used several quality control measures, we acknowledge that there could be inconsistencies with collected worker attributes and questionnaire outcomes.

Conclusion

This work investigates how crowd workers annotate misogynistic and sexist hate speech. Through a MTurk study with 109 workers, we report a significant relationship between hate speech annotation outcomes and workers' political affiliation, moral integrity, benevolent sexist attitudes and personality traits. Furthermore, through interviews with nine workers, we provide detailed insights into how crowd workers make sense of hate speech. We also analyse conditions, structures, and standards that shape the sense-making process, and at what stages crowd workers make their labelling decisions during the annotation process. Overall, our findings inform design considerations for avoiding undesired biases in crowdsourcing subjective judgements, with the potential to enable more robust and fair automated moderation tools and thereby engender safer online spaces.

Acknowledgments

This research was conducted by the ARC Centre of Excellence for Automated Decision-Making and Society (CE200100005), and funded by the Australian Government through the Australian Research Council. We thank the participants for their valuable contributions.

References

Al Kuwatly, H.; Wich, M.; and Groh, G. 2020. Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 184–190. Association for Computational Linguistics.

Alagarai Sampath, H.; Rajeshuni, R.; and Indurkhya, B. 2014. Cognitively Inspired Task Design To Improve User Performance on Crowdsourcing Platforms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, 3665–3674. New York, NY, USA: ACM. ISBN 9781450324731.

Alkhatib, A.; Bernstein, M. S.; and Levi, M. 2017. Examining Crowd Work and Gig Work Through the Historical Lens of Piecework. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, 4599–4616. New York, NY, USA: ACM.

Aroyo, L.; Dixon, L.; Thain, N.; Redfield, O.; and Rosen, R. 2019. Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, 1100–1105. New York, NY, USA: ACM. ISBN 9781450366755.

Aroyo, L.; Lease, M.; Paritosh, P.; and Schaekermann, M. 2022. Data Excellence for AI: Why Should You Care? *Interactions*, 29(2): 66–69.

Barbosa, N. M.; and Chen, M. 2019. Rehumanized Crowdsourcing: A Labeling Framework Addressing Bias and Ethics in Machine Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–12. New York, NY, USA: ACM. ISBN 9781450359702.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623. New York, NY, USA: ACM. ISBN 9781450383097.

Berg, J. 2015. Income Security in the On-Demand Economy: Findings and Policy Lessons From a Survey of Crowdworkers. *Comp. Lab. L. & Pol'y J.*

Berg, J.; and Rani, U. 2021. Working Conditions, Geography and Gender in Global Crowdwork. In *Work and Labour Relations in Global Platform Capitalism*. Edward Elgar Publishing.

Black, J. E.; and Reynolds, W. M. 2016. Development, Reliability, and Validity of the Moral Identity Questionnaire. *Pers. Individ. Dif.*, 97: 120–129.

Borkan, D.; Dixon, L.; Sorensen, J.; Thain, N.; and Vasserman, L. 2019. Nuanced Metrics for Measuring Unintended Bias With Real Data for Text Classification. In *Companion Proceedings of The 2019 World Wide Web Conference*. New York, NY, USA: ACM. ISBN 9781450366755.

Braun, V.; and Clarke, V. 2006. Using Thematic Analysis in Psychology. *Qual. Res. Psychol.*, 3(2): 77–101.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *ICWSM*, 11(1): 512–515.

De Silva, A. 2020. Addressing the Vilification of Women: A Functional Theory of Harm and Implications for Law. *Melbourne University Law Review*, 43(3).

Dosono, B.; and Semaan, B. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, number Paper 142 in CHI '19, 1–13. New York, NY, USA: ACM. ISBN 9781450359702.

Elbanna, A.; and Idowu, A. 2022. Crowdwork, Digital Liminality and the Enactment of Culturally Recognised Alternatives to Western Precarity: Beyond Epistemological Terra Nullius. *European Journal of Information Systems*, 31(1): 128–144.

Felstiner, A. 2011. Working the Crowd: Employment and Labor Law in the Crowdsourcing Industry. *Berkeley Journal of Employment and Labor Law*, 32(1): 143–203.

Fort, K.; Adda, G.; and Cohen, K. B. 2011. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Comput. Linguist. Assoc. Comput. Linguist.*, 37(2): 413–420.

Fortuna, P.; and Nunes, S. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.*, 51(4): 1–30.

Fortuna, P.; Soler, J.; and Wanner, L. 2020. Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 6786–6794. Marseille, France: European Language Resources Association.

Gadiraju, U.; Checco, A.; Gupta, N.; and Demartini, G. 2017. Modus Operandi of Crowd Workers: The Invisible Role of Microtask Work Environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3): 1–29.

Gadiraju, U.; Yang, J.; and Bozzon, A. 2017. Clarity is a Worthwhile Quality: On the Role of Task Clarity in Microtask Crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, 5–14. New York, NY, USA: ACM. ISBN 9781450347082.

Gilbert, S. A. 2020. "I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1): 1–27.

Glick, P.; and Fiske, S. T. 1996. The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *J. Pers. Soc. Psychol.*, 70(3): 491–512.

Goldberg, L. R. 1992. The Development of Markers for the Big-Five Factor Structure. *Psychol. Assess.*, 4(1): 26–42.

Gordon, M. L.; Lam, M. S.; Park, J. S.; Patel, K.; Hancock, J.; Hashimoto, T.; and Bernstein, M. S. 2022. Jury Learning: Integrating Dissenting Voices Into Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM.

Goyal, N.; Kivlichan, I. D.; Rosen, R.; and Vasserman, L. 2022. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. *Proc. ACM Hum. Comput. Interact.*, 6(CSCW2): 1–28.

Gray, M. L.; and Suri, S. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Eamon Dolan Books.

Gray, M. L.; Suri, S.; Ali, S. S.; and Kulkarni, D. 2016. The Crowd is a Collaborative Network. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, 134–147. New York, NY, USA: ACM. ISBN 9781450335928.

Guest, E.; Vidgen, B.; Mittos, A.; Sastry, N.; Tyson, G.; and Margetts, H. 2021. An Expert Annotated Dataset for the Detection of Online Misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 1336–1350. ACL.

Hayes, A. F.; and Krippendorff, K. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Commun. Methods Meas.*, 1(1): 77–89.

Hettiachchi, D.; Sanderson, M.; Goncalves, J.; Hosio, S.; Kazai, G.; Lease, M.; Schaekermann, M.; and Yilmaz, E. 2021. Investigating and Mitigating Biases in Crowdsourced Data. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '21 Companion)*. ACM.

Hettiachchi, D.; van Berkel, N.; Kostakos, V.; and Goncalves, J. 2020. CrowdCog: A Cognitive Skill based System for Heterogeneous Task Assignment and Recommendation in Crowdsourcing. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2): 1–22.

Holton, J. A. 2007. The Coding Process and Its Challenges. *The Sage Handbook of Grounded Theory*, 3: 265–289.

Hube, C.; Fetahu, B.; and Gadiraju, U. 2019. Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–12. New York, NY, USA: ACM. ISBN 9781450359702.

Ipeirotis, P. G. 2010. Demographics of mechanical turk. *CeDER Working Papers*.

Irani, L. C.; and Silberman, M. S. 2013. Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, 611–620. New York, NY, USA: ACM. ISBN 9781450318990.

Kazai, G.; Kamps, J.; and Milic-Frayling, N. 2012. The Face of Quality in Crowdsourcing Relevance Labels: Demographics, Personality and Labeling Accuracy. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, 2583–2586. New York, NY, USA: ACM. ISBN 9781450311564.

Keswani, V.; Lease, M.; and Kenthapadi, K. 2021. Towards Unbiased and Accurate Deferral to Multiple Experts. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, 154–165.

Kocoń, J.; Figas, A.; Gruza, M.; Puchalska, D.; Kajdanowicz, T.; and Kazienko, P. 2021. Offensive, Aggressive, and Hate Speech Analysis: From Data-Centric to Human-Centered Approach. *Inf. Process. Manag.*, 58(5): 102643.

Kumai, K.; Matsubara, M.; Shiraishi, Y.; Wakatsuki, D.; Zhang, J.; Shionome, T.; Kitagawa, H.; and Morishima, A. 2018. Skill-and-Stress-Aware Assignment of Crowd-Worker Groups to Task Streams. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6 of *HCOMP*, 88–97.

Kumar, D.; Kelley, P. G.; Consolvo, S.; Mason, J.; Bursztein, E.; Durumeric, Z.; Thomas, K.; and Bailey, M. 2021. Designing Toxic Content Classification for a Diversity of Perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, 299–318.

Lykourentzou, I.; Antoniou, A.; Naudet, Y.; and Dow, S. P. 2016. Personality Matters: Balancing for Personality Types Leads to Better Outcomes for Crowd Teams. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, 260–273. New York, NY, USA: ACM. ISBN 9781450335928.

MacAvaney, S.; Yao, H.-R.; Yang, E.; Russell, K.; Goharian, N.; and Frieder, O. 2019. Hate Speech Detection: Challenges and Solutions. *PLoS One*, 14(8): e0221152.

Marshall, C. C.; and Shipman, F. M. 2013. Experiences Surveying the Crowd: Reflections on Methods, Participation, and Reliability. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, 234–243. New York, NY, USA: ACM. ISBN 9781450318891.

Meta. 2023. Hate speech. https://transparency.fb.com/de-de/policies/community-standards/hate-speech/. Accessed: 2023-3-10.

Miceli, M.; Schuessler, M.; and Yang, T. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2): 1–25.

Naderi, B. 2018. Who are the Crowdworkers? In *Motivation of Workers on Microtask Crowdsourcing Platforms*, 17–27. Springer.

Nguyen, A.; Halpern, M.; Wallace, B.; and Lease, M. 2016. Probabilistic Modeling for Crowdsourcing Partially-Subjective Ratings. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4 of *HCOMP '16*, 149–158.

Poletto, F.; Basile, V.; Bosco, C.; Patti, V.; and Stranisci, M. 2019. Annotating Hate Speech: Three Schemes at Comparison. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*.

Poletto, F.; Basile, V.; Sanguinetti, M.; Bosco, C.; and Patti, V. 2021. Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review. *Language Resources and Evaluation*, 55(2): 477–523.

Rahman, M. M.; Balakrishnan, D.; Murthy, D.; Kutlu, M.; and Lease, M. 2021. An Information Retrieval Approach to Building Datasets for Hate Speech Detection. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*.

Rea, S. C.; Kleeman, H.; Zhu, Q.; Gilbert, B.; and Yue, C. 2020. Crowdsourcing as a Tool for Research: Methodological, Fair, and Political Considerations. *Bull. Sci. Technol. Soc.*, 40(3-4): 40–53.

Roberts, S. T. 2019. *Behind the Screen*. Yale University Press.

Ross, J.; Irani, L.; Silberman, M. S.; Zaldivar, A.; and Tomlinson, B. 2010. Who Are the Crowdworkers? Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, 2863–2872. New York, NY, USA: ACM. ISBN 9781605589305.

Salminen, J.; Almerekhi, H.; Milenković, M.; Jung, S.-G.; An, J.; Kwak, H.; and Jansen, B. J. 2018a. Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. In *Twelfth International AAAI Conference on Web and Social Media*.

Salminen, J.; Veronesi, F.; Almerekhi, H.; Jung, S.-G.; and Jansen, B. J. 2018b. Online Hate Interpretation Varies by Country, But More by Individual: A Statistical Analysis Using Crowdsourced Ratings. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 88–94.

Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 2021. "Everyone Wants To Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, 1–15. ACM New York, NY, USA.

Sang, Y.; and Stanton, J. 2022. The Origin and Value of Disagreement Among Data Labelers: A Case Study of Individual Differences in Hate Speech Annotation. In *Information for a Better World: Shaping the Global Future*, 425–444. Springer International Publishing.

Sap, M.; Swayamdipta, S.; Vianna, L.; Zhou, X.; Choi, Y.; and Smith, N. A. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5884–5906. Seattle, United States: Association for Computational Linguistics.

Schmidt, A.; and Wiegand, M. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10. Valencia, Spain: Association for Computational Linguistics.

Sen, S.; Giesel, M. E.; Gold, R.; Hillmann, B.; Lesicko, M.; Naden, S.; Russell, J.; Wang, Z. k.; and Hecht, B. 2015. Turkers, Scholars, "Arafat" and "Peace": Cultural Communities and Algorithmic Gold Standards. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, 826–838. New York, NY, USA: ACM. ISBN 9781450329224.

Singer, Y.; and Mittal, M. 2013. Pricing Mechanisms for Crowdsourcing Markets. In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, 1157–1166. New York, NY, USA: ACM. ISBN 9781450320351.

Smith, E. M.; Hall, M.; Kambadur, M.; Presani, E.; and Williams, A. 2022. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9180–9211. Association for Computational Linguistics.

Steiger, M.; Bharucha, T. J.; Venkatagiri, S.; Riedl, M. J.; and Lease, M. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, CHI '21, 1–14. ACM New York, NY, USA.

Twitter. 2023. Hateful Conduct. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy. Accessed: 2023-3-10.

Vallas, S.; and Schor, J. B. 2020. What Do Platforms Do? Understanding the Gig Economy. *Annu. Rev. Sociol.*, 46(1): 273–294.

Vidgen, B.; and Derczynski, L. 2020. Directions in Abusive Language Training Data, a Systematic Review: Garbage In, Garbage Out. *PLoS One*, 15(12): e0243300.

Wiegand, M.; Ruppenhofer, J.; and Kleinbauer, T. 2019. Detection of Abusive Language: The Problem of Biased Datasets. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1*, 602–608. ACL.

Williams, A. C.; Mark, G.; Milland, K.; Lank, E.; and Law, E. 2019. The Perpetual Work Life of Crowdworkers: How Tooling Practices Increase Fragmentation in Crowdwork. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW): 1–28.

Wohn, D. Y. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–13. New York, NY, USA: ACM. ISBN 9781450359702.

Wojatzki, M.; Horsmann, T.; Gold, D.; and Zesch, T. 2018. Do Women Perceive Hate Differently: Examining the Relationship Between Hate Speech, Gender, and Agreement. In *14th Conference on Natural Language Processing*, KONVENS '18.