Open Research Online

The Open University's repository of research publications
and other research outputs

Understanding Misogynoir: A Study of Annotators' Perspectives

Conference or Workshop Item

How to cite:

Kwarteng, Joseph; Farrell, Tracie; Third, Aisling; Burel, Gregoire and Fernandez, Miriam (2023). Understanding Misogynoir: A Study of Annotators' Perspectives. In: 15th ACM Web Science Conference 2023 (WebSci '23), Association for Computing Machinery, New York, NY, United States.

For guidance on citations see FAQs.

© 2023 The Authors

https://creativecommons.org/licenses/by-nc-nd/4.0/

Version: Version of Record

Link(s) to article on publisher's website:
http://dx.doi.org/doi:10.1145/3578503.3583612

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data policy on reuse of materials please consult the policies page.

oro.open.ac.uk

Understanding Misogynoir: A Study of Annotators' Perspectives

Joseph Kwarteng*
The Open University
Milton Keynes, United Kingdom

Tracie Farrell
The Open University
Milton Keynes, United Kingdom

Aisling Third
The Open University
Milton Keynes, United Kingdom

Grégoire Burel
The Open University, UK
Milton Keynes, United Kingdom

Miriam Fernandez
The Open University, UK
Milton Keynes, United Kingdom

ABSTRACT

"Misogynoir" is the anti-Black racist misogyny experienced by Black women, which is characterised by components of both racism and sexism. Misogynoir is challenging to detect due to its inherent subjectivity and its intersectional nature, and people's opinions and interpretations of such hate might vary, which adds to the challenges of understanding it. In this paper, we explored how and some potential why's different annotator characteristics influence how they interpret and annotate a dataset for potential cases of Misogynoir and Allyship. We sampled tweets containing public responses to self-reported misogynoir cases by four prominent Black women in technology, designed an online annotation task study, and recruited annotators of diverse ethnicities and genders from the Prolific crowdsourcing platform. We found that participants' sources of evidence in judging and interpreting content for potential cases of Misogynoir and Allyship, even in circumstances where they all agree on a prospective label, vary across different factors, such as different ethnicity, lived experiences and gender. In addition, we present a variety of plausible interpretations influenced by the various annotators' characteristics. This study demonstrates the relevance of different annotator perspectives and content comprehension in hate speech and the need for further efforts to understand intersectional hate better.

CCS CONCEPTS

• **Social and professional topics** → **Race and ethnicity**.

KEYWORDS

intersectional hate, misogynoir, hate speech detection, datasets annotation, crowdsourcing

ACM Reference Format:
Joseph Kwarteng, Tracie Farrell, Aisling Third, Grégoire Burel, and Miriam Fernandez. 2023. Understanding Misogynoir: A Study of Annotators' Perspectives. In *15th ACM Web Science Conference 2023 (WebSci '23), April 30–May 01, 2023, Evanston, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3578503.3583612

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WebSci '23, April 30–May 01, 2023, Evanston, TX, USA
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0089-7/23/04…$15.00
https://doi.org/10.1145/3578503.3583612

1 INTRODUCTION

"Misogynoir", the anti-Black racist misogyny that Black women experience, which is characterised by both racist and sexist components [18], is a form of intersectional hate. The term was coined in 2008 by Black feminist scholar Moya Bailey, and it focuses on Black women as victims and targets of race and gender-based oppression, since they are vulnerable to both. The stereotype that portrays "Black women as angry", particularly when they speak up, with the intention of undermining what is being said, is an example of Misogynoir [1]. It is challenging to detect due to its inherent complexity and its intersectional nature, as people's opinions of such hateful content might differ based on their demographics and prior experiences [3].

Existing works show how online manifestations of Misogynoir are not easily identified by existing hate speech detection systems. In particular, [6, 12, 16] have shown how the use of computational systems, such as Perspective API [2], HateSonar [3], or Detoxify [4], which are commonly used to automatically detect toxic, abusive, and offensive language, are not as effective when used to identify intersectional forms of hate such as Misogynoir. As they are not sensitive to context, which is a critical component of detecting misogynoir.

The development of automatic systems for detecting forms of hate typically relies on the creation of human-annotated datasets from which these systems can learn. These datasets typically contain statements or phrases that human-annotators will label as hateful or non-hateful. However, annotator characteristics do matter when understanding and interpreting hate [4, 10, 15]. Factors such as being a victim of abuse, ethnicity, racial beliefs, context, gender, and lived experiences may significantly influence people's perceptions and interpretations of hate speech and their labelling behaviour. Likewise, some of these factors may be important in how human-annotators would label potential instances of misogynoir.

In this paper, our focus is understanding how these different human perspectives influence the interpretation of Misogynoir in annotation exercises. In particular, we aim to understand *RQ: What are the differences in how Black women annotate misogynoir in comparison with other groups?*. We hypothesise that Black women might rely more on personal experiences and contextual features (such as; the tweet itself, the dialogue that a given tweet is a part of, other user bio details etc.) in providing a justification for their labelling choices.

[1] https://www.theguardian.com/lifeandstyle/2015/oct/05/what-is-misogynoir
[2] https://perspectiveapi.com/
[3] https://github.com/Hironsan/HateSonar
[4] https://github.com/unitaryai/detoxify

With this purpose, we have designed an annotation study where 80 annotators from the Prolific crowdsourcing platform were primarily asked to annotate potential instances of Misogynoir and Allyship, as an oppositional category to Misogynoir. The content given to annotators was 30 posts, collected from the Twitter social networking site, from four identified misogynoir cases in the Tech sector involving four prominent Black women (see Section 3.1). Annotators were recruited based on their ethnicity and gender, considering four groups of 20 annotators (Black Women, Black Men, White Women and White Men). They were asked to provide: (i) justifications for their choice, (ii) the types of evidence (e.g. the tweet itself, the broader conversation around the tweet, personal experience, etc.) used to interpret the content and, (iii) their level of confidence in their annotation.

Our analysis revealed that even in circumstances where all annotator groups agree on a prospective label, annotators of different ethnicity and gender consider and provide different interpretations and sources of evidence for their arguments, and confidence levels vary between annotator groups. In our study, Black annotators were more likely to be confident in their annotations. On the other hand, Black women annotators may label tweets as possible instances of Misogynoir even if they are uncertain. This may potentially be due to the fact that they have experienced this type of discrimination and are, therefore, more likely to recognise it [4].

Our contributions can be summarised as follows:

- A novel approach for understanding Misogynoir centred on the annotation of the content by four distinct demographic groups, including Black women, Black men, White women, and White men.
- An analysis of how research participants with different ethnicity and gender annotate and interpret Misogynoir and Allyship, contributing to the advancement of the understanding of knowledge around intersectionality in online hate.

The rest of the paper is structured as follows. Section 2 describes relevant related work. Section 3 describes the data annotation process. Section 4 describes our analysis pipeline for the study. Results of the analysis are presented in Section 5. Discussions and conclusions are presented in sections 6 and 7, respectively. [5]

2 RELATED WORK

Studying hate speech online is not just about content moderation; it is about building a better understanding of the phenomenon so that we can address it as a society. For the most part, computational approaches toward studying hate speech online have largely addressed either single-axis hate speech (based on one characteristic) or, more generally, toxic or abusive and offensive language [1, 2, 8, 9, 19]. Intersectional forms of hate, directed at those with multiple identities marginalised in society, are more complicated to identify and understand online. For instance, prior research that investigated Misogynoir using a lexical-based approach revealed that it was ineffective due to the contextual and nuanced nature of the hate [5].

Previous studies have demonstrated that factors such as being a victim of abuse, ethnicity, racial beliefs, context, lived experiences, and gender may significantly influence people's perceptions and interpretation of hate speech and their labelling behaviour [4, 10, 15]. In terms of victims and targets, people who have directly experienced abuse in the past are more likely to label a random statement as toxic, as well as groups historically at risk of abuse [4]. Roussos and Dovidio [14], for example, studied how anti-Black prejudice may impact perceived differences in free speech protections (FSPs) for hurtful behaviours directed at Black or White individuals. In a first study, they presented a short vignette detailing an incident in which the perpetrator explicitly expressed negative group-based sentiments and asked whether the act and the corresponding comment were about White or Black people. In a second study, they modified the stimulus material and focused solely on White participants, exploring further factors that could account for previous responses of low anti-Black bias participants. Across the two investigations, they discovered that when the hate target was Black rather than White, those with more anti-Black prejudice were more likely to claim freedom of expression rights and viewed an incident as less of a hate crime. In contrast, persons with low levels of prejudice viewed the conduct as less protected by freedom of speech rights and as a hate crime. Sap and Swayamdipta [17] also examined annotator attitudes and investigated the *who*; who annotates toxic, the *why*; why they annotate as toxic and *what*; what is being rated, concluding that a strong correlation exists between annotator identities and beliefs and their ratings of toxicity. In addition to the *what*, Sap et al. [16] examined different tweets and discovered that African American English (AAE) tweets are twice as likely to receive a higher toxicity rating than others, which indicates potential risks and the importance of comprehending the *what* (content) and the language being communicated. Ross et al. [13] discovered that providing annotator rules with the definition of hateful behaviour to annotators was insufficient since annotators partially aligned their beliefs with the definitions. This shows the importance of having a diverse pool of annotators to understand their perceptions and biases better.

These studies all show that annotators' perspectives do make a difference when assessing hateful and toxic content. In our work, we investigate how the different annotator groups understand, interpret and justify potential instances of Misogynoir as a way to understand the phenomenon better. We ask the different annotators to justify (the *why*) and to name the source of their justification (i.e. the evidence used).

3 DATA ANNOTATION

In this section, we describe how we have generated the annotated dataset used for this study. We describe the Twitter data used for the annotation (Section 3.1), The annotation study design (Section 3.2), and the annotator recruitment process (Section 3.3).

3.1 Dataset

To set up our annotation study, we purposefully sampled and used 30 tweets from a previously collected dataset [5]. The tweets in this dataset were gathered by collecting the public response on Twitter towards the self-reported experiences of misogynoir of

[5]The code and the data used for the analysis will be made is publicly available under https://github.com/kwartengj/WebSci23_Misogynoir. For the annotated tweets only tweet IDs will be provided following Twitter's Terms of Service and Publishing Guidelines.

four high-profile Black women working in the technology sector. Tweets were collected from the time they shared their stories until January 2021. The data collection process used the individuals' Twitter handles and their displayed names as keywords. Since we required annotators to provide a wide range of information for each of the annotated tweets, i.e., the annotation, their reasoning for the labels, their source of justification, and their level of confidence in their annotation, we selected fewer tweets to obtain qualitatively rich data.

3.2 Study Design and Task

We designed an online survey to capture annotators' rationale when assessing the previously selected 30 tweets. Each annotator was asked to categorise the 30 tweets into four mutually exclusive categories: Misogynoir (M), Allyship (A), Unclear (U), or None of the Above (NA). Brief explanations were provided for each of the categories:

- Misogynoir (M): a tweet may be labelled as a possible instance of misogynoir if it expresses anti-Black racist misogyny (hatred) against Black women, which has both sexist and racist features.
- Allyship (A): a tweet might be considered a possible instance of allyship if it shows any form of support or solidarity to Black women.
- Unclear (U): a tweet may be labelled unclear if the annotator is uncertain about its meaning and cannot determine if it is misogynoir or allyship.
- None of the above (NA): a tweet may be labelled none of the above if the annotator is certain about its meaning and can determine that it is neither misogynoir nor allyship nor unclear.

Participants were also asked to offer a justification for their classification as well as to indicate their level of confidence in the labels they chose (Completely Confident, Fairly Confident, Not Confident at All, Slightly Confident, and Somewhat Confident). In the final step, annotators were asked to identify their source of justification, i.e., the factors that influenced their judgment of the tweets. We presented them with a few examples and gave them the option to add other ones they utilised (see Figure 1).

3.3 Annotator Recruitment

Annotators were recruited via Prolific[6], a popular online crowdsourcing platform. It provides a range of demographic detail about its participant pool on its website, which researchers can also use to screen Annotators and recruit for their various studies.[7] Previous research found that Prolific participants are more diverse, less exposed to standard research tasks and therefore (potentially) more sincere in completing tasks [11].

Given the scale of this task and the need to recruit a diverse sample of workers across Prolific, it was necessary to provide a competitive remuneration. After piloting, we decided to pay £10 as compensation, which is slightly more than the platform's minimum wage. It took annotators an average of about 50-60 minutes to complete the task. Annotators were recruited to one of four groups,

[6]https://www.prolific.co/
[7]https://www.prolific.co/about

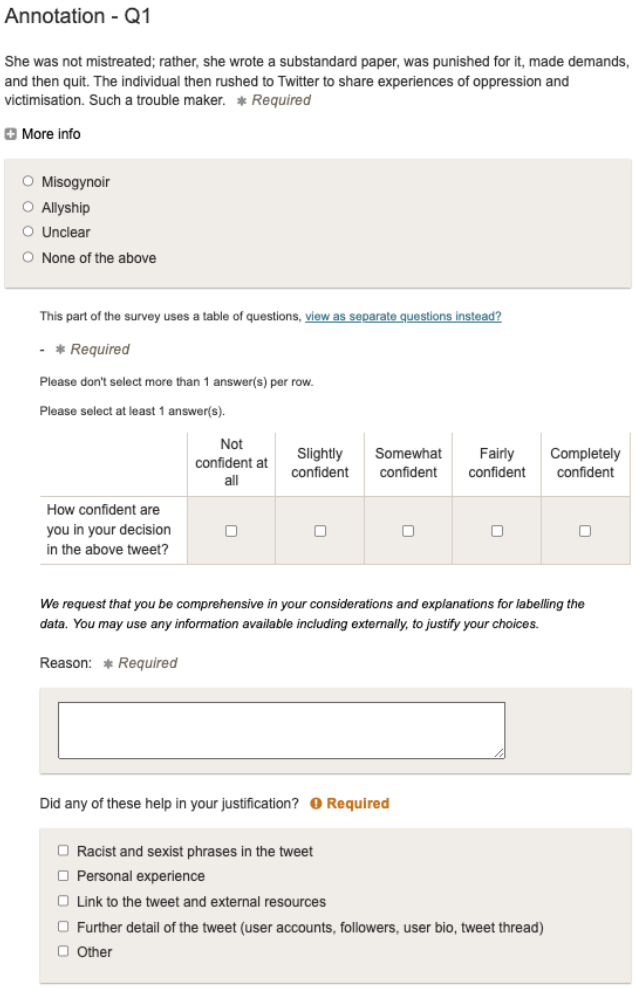

Figure 1: Example of the annotation process for a tweet.

based on their identity as i) Black women, ii) Black men, iii) White women, or iv) White men, and were all above the age of 18. Our target was to sample about 80 Annotators equally distributed across these groups (see Table 1 for Annotators distribution).

We did not set any restrictions in terms of location because we wanted the sample to be diverse in terms of location.

Table 1: Distribution of annotators per each group

Groups	No. of Annotators	Avg. Age
Black women (BW)	20	25.7
Black men (BM)	20	25.6
White women (WW)	20	27.7
White men (WM)	20	27.1

4 ANALYSIS SETUP

To answer the research question, we examined the annotators' label ratings and computed the Fleiss' kappa[8] inter-rater agreement across groups. We then compared the kappa value of Black women to that of the other groups to ascertain the rating difference and to assess if Black women could identify more diverse examples of Misogynoir and Allyship than the annotators in the other groups, i.e. White men and women and Black men.

To grasp the influence of race, gender, and their intersection on the annotator ratings, we classified the annotation into two groups; a group of Race consisting of Blacks (Black women and men) and Whites (White women and men) and a group of Gender consisting of Males (Black men and White men) and Females (Black women and White women). We then calculated the Fleiss' kappa inter-rater agreement between the groups. We further analysed the source of justification (i.e. what annotators considered beneficial in their annotation, e.g. link to the tweet, racist and sexist phrases etc.) and the confidence for their annotations to acquire a better picture of what the different groups thought useful in their annotation task and how confident they were in their annotations.

We further established a metric to evaluate conflicts and agreements per tweet. The metric is based on annotation ratios per category and a predetermined threshold t.

Lets TA be the total number of annotations for a tweet tw_i and M_i, A_i, U_i, NA_i the number of annotations for tw_i that fall into the categories of Misogynoir, Allyship, Unclear, and None of the Above, respectively. We then compute the ratio of annotations per category for each tweet tw_i (i.e., M_i/TA, A_i/TA, U_i/TA and NA_i/TA). We then select the maximum value of these ratios and compare it to the threshold. If the value is more than or equal to the threshold, we then consider that there is an overall agreement over the tweet, and we classified the tweet into the category with the maximum ratio; otherwise, we consider that there is a disagreement over the tweet.

We then looked at the tweets for which the previous step showed that there was a disagreement between annotators and classified them into three distinct annotation categories: bi-disagreements, tri-disagreements, and quad-disagreements. A bi-disagreement means the annotators' judgements are split between two categories, a tri-disagreement between three, and a quad-disagreement between four. For $2 \leq n \leq 4$, a tweet is an n-disagreement if it is not an agreement, not an $n-1$ disagreement (where defined), and there are n categories such that the sum of the annotation ratios for those categories is greater than or equal to t. Essentially, if n categories are collapsed into a single category (by adding their annotation ratios), we ask whether annotators agreed that the tweet is that category (in the same sense of agreement according to threshold t).

We then calculated an "annotation grouping" for each tweet according to the proportion of annotators who have annotated a tweet with a given label using the disagreement and agreement metrics described above with a threshold of 0.75. We chose this threshold empirically because it allows a more rigorous evaluation of disagreement and agreement among annotators per tweet, ensuring that the final annotation grouping is based on a substantial majority consensus rather than just a simple majority or random chance.

These annotation groupings allow us to see how each tweet was annotated by each participant group (see Table 4). Using this table, we can identify tweets where the annotation grouping for Black women is the same or different from the annotation grouping for other participant groups. For example, we can see that tweets 10 and 12 have an annotation grouping of Misogynoir for all participant groups. So there is an agreement between Black women and other annotators. We can also see disagreements and gain an understanding of how meaningful these disagreements are. For example, tweet 6 is contested among Black women primarily between the annotation categories of Misogynoir and Allyship. For Black Men, it is contested between three categories of Misogynoir, Unclear or None of the Above. For White women, the tweet is contested across all categories (Misogynoir, Allyship, Unclear, and None of the Above) and for White Men, the tweet is contested between Misogynoir or None of the Above. For the purposes of our analysis, we are interested in tweets where Black women do not agree with any other groups, tweets in which Black women agree with Black men but not other groups, and tweets in which Black women agree with White women but not other groups.

To be more selective for the scope of this paper, we introduced some additional criteria to identify the tweets we were interested in examining further. As Misogynoir and Allyship are oppositional categories, we are particularly interested in tweets where annotators had difficulty deciding between these two categories (a binary disagreement) or where both Allyship and Misogynoir are present in the annotation grouping (such as M or A or U and M or A or U or NA). However, for the purposes of this paper, we are not yet exploring disagreements that are related to the Unclear and None of the Above annotations when there is no additional disagreement between Misogynoir and Allyship.

We arrived at a selection of tweets 5, 6, 9, 17, 22, and 24 that needed to be examined further (see Table 4). While it was outside the scope of this paper to do a comprehensive qualitative examination of all justifications given for each tweet, we did consider the wording of the tweet, and we were able to identify different interpretations of the tweet based on some of the justifications given by different participant groups that may explain where the disagreement originates.

5 ANALYSIS RESULTS

In this section, we report the results of our quantitative analysis together with our examined tweets analysis to contextualise our findings. We follow our primary research question to guide this analysis: **What are the differences in how Black women annotate this dataset for misogynoir and allyship in comparison with other groups?**

5.1 Differences in annotations by Black women vs other groups:

Figure 2 displays the number of annotations within each category for each of the four groups and demonstrates that the Allyship category is easily recognised across all groups. This may be owing

[8] https://en.wikipedia.org/wiki/Fleiss%27_kappa

to the fact that the majority of tweets labelled as allyship by annotators had words and concepts such as "support", "thank you", and "solidarity" that made it easy to detect.

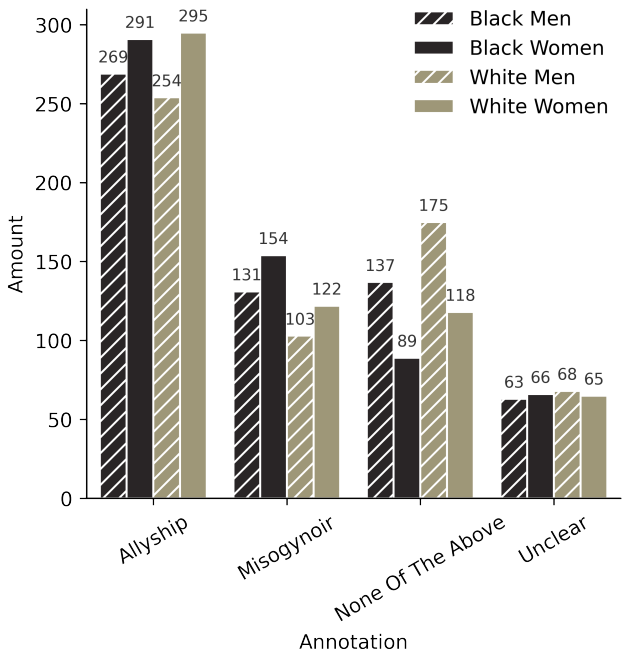

Figure 2: Distribution of annotation categories per each group

Black women had 291 potential cases of Allyship annotations, whereas White women had 295, Black men had 269, and White men had 254. In terms of Misogynoir, we tend to see Black women identifying more instances of tweets as a potential case of Misogynoir with 154 annotations, followed by Black men with 131 annotations, White women with 122 and White men with 103. The reason for this could be because Black women experience racism and misogyny and could potentially identify instances of messages with such elements, Black men experience racism and could also potentially identify instances of messages with racist elements, White women experience sexism and could potentially identify instances of sexism in these messages, whereas White men do not experience this type of hate and annotating such messages could be a potential challenge for them. Thus, Misogynoir by definition is a type of hate that affect Black women and is characterised by racism and sexism. Since White men do not experience any of the components of this hate (i.e racism or sexism), it could potentially be a challenge for them to annotate.

5.2 Difference in Inter-annotator agreement across Race and Gender:

Table 2 shows the inter-annotator agreement for each category across the four groups. As we can see from the table, Black women had the highest overall Fleiss' kappa inter-annotator agreement score of (0.32) compared to White men's (0.305), Black men's (0.257) and White women's (0.261). These scores range from (0.21) to (0.40), indicating "fair agreement" [7]. Category-wise, Black women had the highest agreement score for the Allyship category, (0.483), indicating "moderate agreement", whereas White women had the highest agreement score for the Misogynoir category, 0.411, indicating "moderate agreement". Notably, this does not imply that White women can identify more instances of Misogynoir tweets than Black women, but rather that they had a greater consensus on the tweets they qualified as such.

Table 2: Inter-annotator agreement per category for each of the four groups

Categories	BW	BM	WM	WW
Allyship	**0.483**	0.401	0.455	0.343
Misogynoir	0.374	0.303	0.366	**0.411**
None of the Above	0.115	0.119	0.188	0.101
Unclear	0.068	0.051	0.094	0.062
Overall Kappa	**0.32**	**0.257**	**0.305**	**0.261**

In table 3, we present inter-annotator agreements based on gender and race independently. The table reveals that Blacks and Women obtained an overall greater degree of agreement, (0.29) and (0.287), respectively than Whites and Men, (0.27) and (0.274), within the fair agreement range. In terms of race, Blacks had the highest kappa value, (0.443) ("moderate agreement"), compared to Whites, who had a score of 0.396 ("fair agreement"). Comparing the kappa value for gender, the Women group scored (0.387) ("fair agreement") vs (0.334) ("fair agreement") for the Men group. This result indicates that Black people and women had a greater level of agreement than White people and males among the tweets tagged as Misogynoir and Allyship. This may be due to the fact that Black people and women experience distinct forms of prejudice that White people and men rarely encounter; it may be easier for them to recognise such prejudice when they meet it.

Table 3: Inter-Annotator agreements per annotation categories for the different Race & Gender groups

	Race		Gender	
Categories	Blacks	Whites	Women	Men
Allyship	**0.443**	0.396	0.410	**0.419**
Misogynoir	0.340	**0.390**	**0.387**	0.334
None of the Above	0.118	0.138	0.093	0.144
Unclear	0.071	0.08	0.075	0.067
Overall Kappa	**0.29**	**0.27**	**0.287**	**0.274**

5.3 Difference in annotator confidence per categories between Black women vs other groups:

In Figure 3, we present the average confidence rating of annotations per category for each of the four groups. In the figure, we can observe that Black women and Black men are more confident in their annotations for the category Allyship than White women and

men, as they had a higher average confidence rating. In terms of Misogynoir, we found Black men scoring higher averagely compared to Black women, White men and White women. However, in table 4 and Figure 2, we can see that Black women also annotated more potential cases of Misogynoir tweets; is this confidence level because they were more likely to label tweets as potential cases of Misogynoir, even if they were not completely certain?

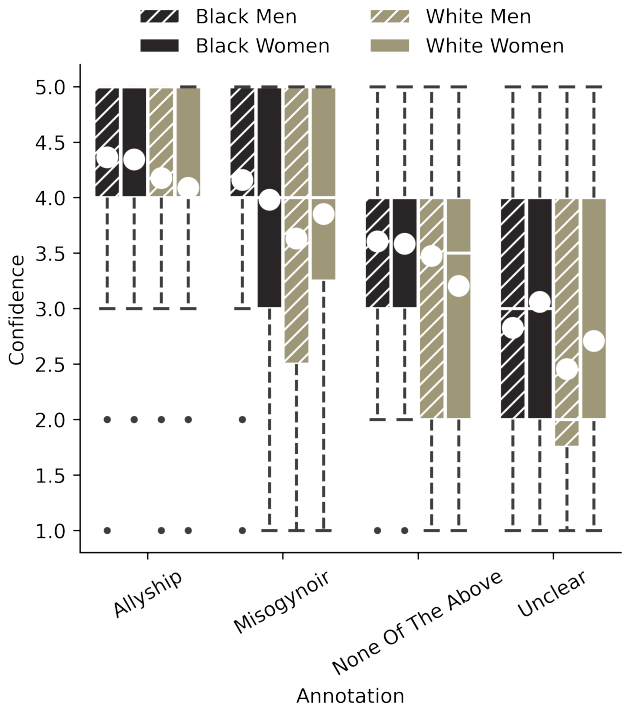

Figure 3: Average confidence of annotation categories across the annotator groups; white dot = mean

5.4 Differences in annotator sources of justification per categories between Black women vs other groups:

In examining what informational resources annotators believed to have assisted them in annotating the tweets, Figure 4 presents the percentages of informational resources used per each group. Note that some information resources may be combined for a single tweet. For example, annotators may use links, further details and personal experience as an information resource that assists them in annotating a single tweet. Figure 4 reveals that all annotation groups deemed the tweet's link to be a helpful source of information when annotating tweets. White men (0.70) considered it more beneficial, followed by Black men (0.65), Black women (0.54) and White women (0.47). In terms of further details (i.e. following up on the user's account and bio, followers, tweet thread, etc.), Black women and White men find it equally useful, with (0.34) from each group indicating that it aided them in their annotation, compared to (0.37) from White women, who found it the most useful, and (0.26) from Black men.

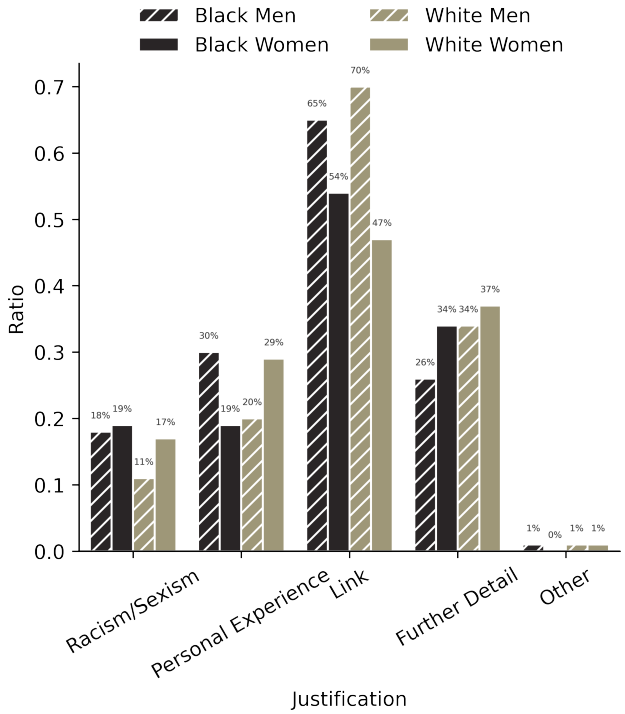

Figure 4: Percentages of annotator's source of evidence for their justification for labelling a tweet for each group

Black women consider racism/sexism (i.e., racist and sexist words or concepts) most helpful, with (0.19) finding it most beneficial, followed by Black men (0.18), White women (0.17) and White men (0.11). This implies that, at least in some instances, all groups considered the context of the tweet beneficial and also noticed racist and sexist phrases used. In terms of personal experience (i.e. drawing from earlier encounters or experiences), the intriguing result is that groups other than Black men relied more on personal experiences in their annotations than Black women, with (0.30) of Black men considering it as the most useful, followed by (0.29) of White women, (0.20) of White men and then (0.19) of Black women. However, our tweet examination found otherwise, as Black women drew mostly from their personal experiences than groups other than Black women. Although it makes sense for the minority groups, i.e., Black women, Black men, and White women, who have all experienced some sort of hate, to draw from those experiences, it is difficult to identify the specific experiences from which White men drew, potentially classicism and ableism.

5.5 For which tweets do we see Black women agree with each other and with other groups?:

From table 4, we see that tweets 10 and 12 is a case where Black women agree with each other and with other groups. Tweet 10

Table 4: Annotation groupings and their tweet IDs by participant group

	Participant group			
Annotation groupings	Black women	Black men	White women	White men
M	[7,10,12]	[7,10,12]	[7,10,12]	[10,12]
A	[11,13,14,15,16,18,19,20,25]	[11,13,14,15,18,19,20,26]	[9,11,13,14,15,16,18,19,20]	[11,13,14,15,18,19,20,25]
U	-	-	-	-
NA	-	-	-	-
M or A	[6, 28]	[28, 29]	[5,28]	[28]
M or U	-	-	-	-
M or NA	[4, 8, 24]	[4, 8]	[4,8]	[2,6,7,8]
A or U	[17]	-	-	[16]
A or NA	[27,30]	[9,16,17,22,25,27,30]	[17,22,25,26,27,29,30]	[9,17,21,22,24,26,27,30]
U or NA	-	[24]	-	-
M or A or U	[2,9]	[2]	[2]	-
M or A or NA	-	[3,5]	-	[5]
M or U or NA	[1,5]	[1, 6, 23]	[1]	[1,29]
A or U or NA	[21,23]	[21]	[21,23,24]	[3,23]
M or A or U or NA	[3, 29, 22, 26]	-	[3, 6]	[4]

reads ***"It appears as though you would never be pleased unless Google was entirely black and racist in the opposite direction."*** and Tweet 12 reads ***"PUT AN END TO THE ANGRY BLACK WOMAN STEREOTYPE ALREADY"***. Given that we informed annotators that misogynoir is hatred directed specifically towards Black women and the fact the tweets expressly insult Black women, might have made it easy for the annotators. Misogynoir is the only plausible interpretation for these tweets. From figure 5, we see the source justification for tweets 10 and 12 all combined. We can see that despite the fact that all groups of annotators included racism/sexism as a source of evidence for their justifications, White women (0.75) and Black men (0.70) considered it much more than White men (0.65) and Black women (0.65). The Female annotators indicated the use of further details as a source of evidence much more with (0.28) of Black women and (0.25) White men compared to (0.15) each for Black men and White men. In addition, Black annotators are more likely to utilise personal experience as a source of evidence than White annotators, with (0.30) of the Black men and (0.20) using personal experience as a source of evidence for their justification, compared to White annotators, with (0.18) of White men and (0.15) of White women.

Tweets 11,13,14,15,18,19, and 20 were identified as one of the cases from our pipeline where Black women agreed with other groups; see table 4. These tweets could be interpreted as showing consensus among the annotator groups that this is a possible instance of Allyship. Figure 6 presents the combined source justification for the tweets. We can observe that link is the most often regarded source of annotator evidence for justification and that it is mostly considered by males, with (0.77) of White men and (0.67) of Black men, compared to (0.57) of Black women and (0.46) of White women for the female group. racism/sexism is less likely to be used as a source of evidence to support a probable case of Allyship since all annotator groups rated it lower than the other sources. This is simply explained by the fact that the definition of Allyship does not contain such terms and phrases.

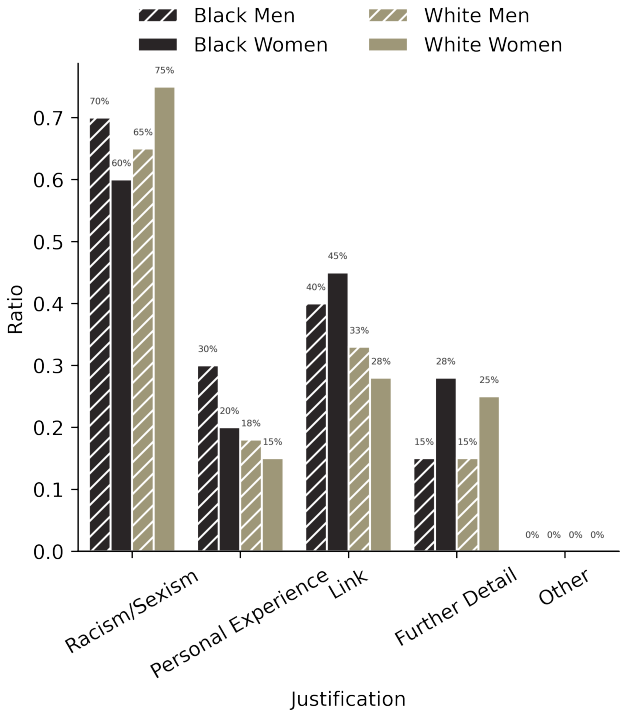

Figure 5: Tweets 10 and 12: sources of justification per annotator groups

From table 4, we can see that tweets 1, 8, 28 and 30 were in the same annotation groupings for all annotator groups and could appear as Black women agreeing with other groups. These tweets are disagreements (see section 4), because no consensus was reached for a potential category. However, because Black women agree with

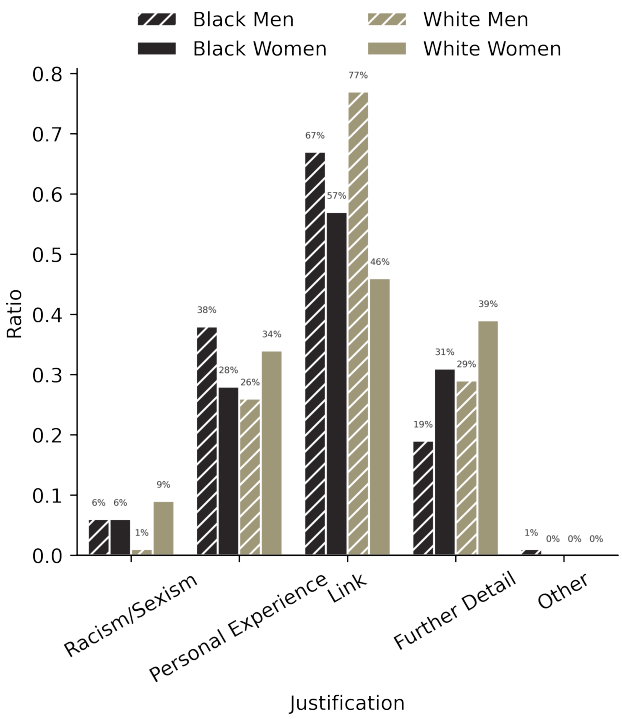

Figure 6: Tweets 11,13,14,15,17,19 and 20: source of justification per annotator groups

other groups on the nature of the disagreement (that the choice is between M and A for example), we did not prioritise these tweets for extended analysis. This is something we will explore in future work.

5.6 For which tweets do we see Black women disagree with all other groups?:

Tweet 5 reads **"I may have erased and reposted it to remedy an error, but I'm unsure. Following certain AI professionals who support you piqued my attention, so I investigated more. I felt that it was unclear. Sending outraged emails to organisations also results in dismissal for white guys (e.g. me). Best of luck."** and it appears to be associated with the labels Misogynoir or Unclear, or None of the Above. One possible interpretation of this tweet is the author's take on the target's actions and possible consequences. For example, one White woman wrote; *"This man says that some actions can happen to other races as well and it doesn't matter if you are a man, or woman, black, or white"*. One Black female annotator also wrote: *"A white man clearly saying he is provoked by a women speaking out and writing 'angry emails' to speak out will provoke him to fire her is very misogynoir."* Another annotator used some additional evidence to check this supposition. The Black female annotator wrote: *"He was using dismissive language in the tweet and is kind of implying that what happened to her can happen to anyone hence him mentioning that "white men" can get fired too. He calls it sending "angry" emails instead of sending emails speaking up for yourself. This comes across as him seeing her as an "angry" black woman because he could have phrased it in a different way instead of using the word angry."*

Other plausible readings imply that the author did not comprehend the circumstance and drew unfounded conclusions. One White male annotator wrote: *"And again, turning a victim into a instigator and a person acting bad. The part talking about "firing for white men" also quite speaks for itself. That's an example of not understanding the problem, but in the same time hitting the minorities."*

Alternatively, some interpretations of the tweets argue that they are ambiguous and maybe neutral but somehow with a racist intent (from the previously deleted tweet). For example, two quotes from White women includes: *"the message is neutral"* and *"The message of the tweet itself is unclear. The Tweeter himself is a right twat in that thread history - first racist, then deletes when gets backslash, then backpedals and makes excuses. Classic behaviour."* Another White male annotator also wrote: *"didnt really understand it but there where some racist things about the white people, that he will provoke to f.e. fire those white men and etc"* This interpretation is not limited to groups who are non-Black. Two quotes from Black female and male annotators include: *"I don't know where they stand"* and *"This tweet is fairly benign and neutral comment, but the context from which it emerges has strong racist undertones (hired because of her race)."*

In Tweet 5, these varying perspectives and interpretations emerge due to the disruption of the context (deleted tweet from the author) and annotators are not able to follow the conversation. In an online environment, this may completely alter the intended meaning of the text, making it more difficult to grasp and also influencing how annotators view and interpret it, as evidenced by our annotators' interpretations.

Tweet 6 was identified as an interesting case from our pipeline, the text reads **"So a Black woman lost her job because she spoke out against racial bias? What proof do you have for your assertion?"**. One possible interpretation of this tweet, which appears to be more closely related to the labels Unclear or None of the Above, is that the author is just seeking clarification. For example, one White male annotator wrote: *"I don't think this could be a misogynoir or allyship tweet, I think that they just look for some clarification on given subject"* Another White female annotator wrote: *"just asking for evidence after someone claimed that The lady was fired after raising issues about racial bias"*. This interpretation is not limited to groups who are not Black women. Two quotes from Black female annotators include *"They want all the facts before being judgemental"* and *"This is a question posed to the original tweet author. It is polite."*

It appears that these justifications take the tweet at face value. Other interpretations include questioning the tweet author's sincerity based on the account details (few followers, new account, prior tweets) and, more generally questioning the tweet author based on existing knowledge of misogynoir. For example, one Black female annotator wrote: *"i chose that this use is Misogynoir. As a black woman when you raise a point, you will always be asked to 'prove it' when most instances it is not something that you can physically display to people as it was her own experience and that should be enough reason"*. Another annotator wrote: *"Again, I find that*

the comment is trying to belittle Timmit's experience by demanding evidence or proof. Why do they feel she is accountable to them?"

Black men's justifications for labelling this tweet as Misogynoir also suggest that the need for evidence is more than simply a formality. For instance, one annotator wrote: *"He does not believe that [Black woman's name] was fired for raising issues about racial bias, they even have doubts that this is true"*. Another annotator used some additional evidence to check this supposition: *"This tweet is a response to another tweet which said that [Black woman's name] was discriminately fired. However this tweet has repeated the statement with a question mark. This could be seen as misogynoir because black women are not taken seriously when an event occurs. Instead the tweeter asks for evidence that it has occured. Following on from the tweet, the original tweeter has said that the evidence is online."*

Understanding the context, sharing the same mastery command of the language, and understanding what words and phrases mean to various individuals may impact how you see and interpret things, as illustrated by Tweet 6. Unfortunately, in an online environment where not everyone shares a language, the same mastery of that language, or the same interpretations of words, this makes everything extremely difficult as people might draw different interpretations which might not entirely be the case.

Tweets 9 and 17 were identified as noteworthy cases from our pipeline and were grouped together since they both have the Allyship label closely near the threshold (See section 4) though they have closely associated with the labels Allyship or None of the Above, or Unclear. Tweet 9 reads **"It sounds as though you accomplished remarkable feats throughout your stay there. However, waiting until after you've been dismissed to discuss such matters casts doubt on your motivations. You're understandably furious and hurt. In any event, I am confident that your voice will continue to advance acceptability."** and tweet 17 reads **"I appreciate your sharing your experience."**

One possible interpretation of tweets 9 and 17 is that the person is just showing support. For example, one Black male annotator wrote: *"Shows a level of solidarity and support"* [tweet 17] and *"constructive criticism here, with support hoping for progress."* [tweet 9]. Another Black women annotator wrote: *"The tweet shows support for the woman. This person showed kindness and thanked the woman for sharing her story."*[tweet 17] and *"In total support"*. [tweet 9] White annotators shared the same interpretation. One White female wrote: *"In my opinion the author of the tweet supports the person but not 100%"* [tweet 9] and *"The tweet is supportive and It shows that she is not alone in exposing this."*[tweet 17]. Another White female also wrote: *"supports the victims"*[tweet 17].

Other interpretations, however, do not detect any discriminating language or any sign of attack or hatred. For instance, one Black male annotator wrote: *"No sexist or racist phrases. THis was a response to a person advocating for women of colour"*[tweet 9]. One White female also wrote: *"The person is calm, doesnt attack her and sees the tweet as a positive action that people will understand her"*.

Alternate interpretations for tweet 9 suggest that the author was only seeking logical connections and questioning everything is acceptable. For example, one White male annotator wrote: *"Again, there are no signs of the person being attacked or defended, but the author is trying to find logical connection between events."*. Another White female annotator also wrote: *"I'm not very confident on this. It is a fairly supportive tweet but still there is a little dig about motives. But honestly, questioning everything is the basis of academia, so this is acceptable."* However, another interpretation finds it confusing as the author supports and questions her motives at the same time. For example, one White female annotator wrote: *"I am slightly confused about this one. On one hand it seems as though the person is a supporter of the girl trying to fight against sexism and racism in the workplace, but at the same time he says he "questions her motives". However the last sentence does mostly indicate support, so I labeled it as allyship."*

The Black female annotators who labelled tweet 9 as Misogynoir suspect that questioning her motives is a common "sarcasm" to silence Black women. Two quotes from Black female annotators include: *"Just because a person does not speak about their problems at work does not mean they are not aware of them. The lady wasn't at that time to share her experience and that is understandable."* and *"This is the sarcasm commonly used when women speak-up, it is usually done to invalidate their feelings. Happening of events is important regardless of anything."* Another annotator used some additional evidence to check this supposition: *"The person in question is being made out to be 'the angry woman' because she is now coming out to speak on the situation. This picture has been painted and happens a lot to black women on a regular."*

Tweet 9 potentially is a clear example of text that could be interpreted as being two things at once, as seen from the different annotator interpretations. Nonetheless, we miss out on these key nuances and interpretations since debates about hate speech online are still rather binary - either it is (i.e., hateful) or it is not (i.e. non-hateful).

Tweet 22 reads **"Former Pinterest workers and five others give examples of the company's inability to address racial discrimination and bigotry in a piece for the Washington Post"**. The tweet appears to be associated with all the labels; one plausible interpretation is that the author posted a news article. For example, one Black female annotator wrote: *"It's just someone sharing information that there are other victims that have spoken up. No direct stance of where they stand"*. Three White male annotators also wrote: *"Again a comment conveying news without taking sides."*, *"Just a tweet of an artice quoting from said article. No show of allyship or misogynoir."* and *"This is only an informative tweet."*

The Black and White female annotators who labelled it Allyship stated that the author shared to bolster the cause and that the author is showing support. For example, two White female annotators wrote: *"The author of the tweet is supportive. They quoted the article about people talking about Pinterest's failure in addressing racial discrimination"* and *"Shows support and gives other similar examples give strenth to the cause"*. Two quotes from Black female annotators include *"i think the fact that they chose to share this information is enough to show their stance on this matter. the user is trying to highlight that fact that she is not alone in this struggle"* and *"The comment makes mention of a situation that is similar to Timmit's in support of her"*. However, this Allyship perspective from the female annotate was not exclusive to women, as we discovered comparable justification within the male group. For example, one Black male annotator wrote; *"the author is using other people's experiences to show support"* and one White male also wrote; *"The author is trying to get people to support by spreading the message"*.

Tweet 22 is a potential example of how, in the online world, others may associate what you share or discuss as something that is significant to you or that you support. These varying interpretations agree that it is a prospective news item, but the author's decision to share it alters its interpretations.

Tweet 24 appears to be another interesting case, the tweet reads **"otherwise, you would wind up with algorithms that create outcomes that you would "want" to see rather than what is actually occurring, and you wouldn't even realise. Even worse, you would be encouraged to believe that you are correct when you are utterly incorrect."** The tweet appears to be associated with the labels Misogynoir or None of the above and has a general lack of clarity. For example, two Black male annotators both wrote: *"I dont understand his statement"*. Another Black female annotator wrote: *"not enough information on tweet as to what exactly they are referring too"*

This interpretation is not limited to the Black group. Two quotes from White male annotators include: *"I don't understand what he says"* and *"can not understand the context of the message."* One White female annotator also wrote: *"This entry is very vague. I don't know what exactly it is about."*

However, alternate interpretations include: doubting the tweet's relevance to the issue and finding no racist or sexist language. For example, one White male annotator wrote: *"That tweet is not related to the black women situation"*. Another White female annotator wrote: *"There is nothing about racism or approving Black women. This annotation touches different topic."*. One Black female annotator wrote: *"The annotation speaks on a totally different topic so it would not be either misogynoir or allyship."*

Black women's justifications for labelling this tweet as misogynoir also contended that the author was disrespectful and invalidated the lived experiences of the Black woman involved. For example, one annotator wrote: *"This person clearly has a problem with the woman. The tweet shows disrespect towards the woman."* Another Black female annotator wrote: *"In my view this comment sounds like it is invalidating her experience. I do stabd to be corrected"*.

Similar to this tweet, other tweets are quite confusing, and it will be necessary to investigate them further.

6 DISCUSSION

Previous research has indicated that diverse life experiences, gender, and ethnicity have a substantial impact on people's labelling behaviour [4, 10, 15]. Our results revealed that people of different races and gender identities consider different pieces of evidence in judging and interpreting content for potential cases of Misogynoir and Allyship. Our study went further to show that, even in circumstances where they all agree on a prospective label, viable interpretations, sources of evidence for their arguments, and confidence ratings vary between annotator groups, which may be viewed as the impact of various life experiences and their influences on their judgments and comprehensions. In addition, our study found that prospective instances of Allyship are more likely to be identified by the various annotators, particularly women, whereas prospective Misogynoir cases are more likely to be identified by Black women and the Black group.

Our qualitative examination of a number of participant justifications revealed that the disruption of the content provided (i.e. annotators not being able to follow or fully comprehend the content or conversation or deleted content from the tweet's author), the understanding of the context (i.e. sharing mastery command of the language and what its phrases mean), the lack of clarity of the content, and the author of the post's legacy data could influence the different perspectives of the annotators from different demographics.

Regarding the inter-annotator agreements across the groups, the findings suggest that Black women reached a higher consensus than the other groups. Also, the women and the Black groups had a higher consensus value than the Men and the White group. This may potentially be interpreted as the fact that Black people and women experience distinct forms of prejudice (for example; misogyny, racism etc.) that White people and men rarely encounter; which may make it clearly identifiable to them as to groups who rarely see it.

Black women and Black men are more confident on average in their annotation compared to White men and White women for both potential cases of Allyship and Misogynoir. However, though Black women were found to have annotated more potential instances of Misogynoir than Black men, their average confidence rating was lower than Black men. A more plausible explanation could be that Black women directly experience this form of abuse and may be more likely to label tweets similar to their experiences as Misogynoir even if they are not entirely confident. Also, on the basis of their experience, it is possible that Black women are more attuned to prospective instances or able to recognise more nuanced examples of potential instances of Misogynoir. This is consistent with the findings that groups who have historically been at risk or who have directly experienced abuse are more likely to label a statement as toxic if it is relevant to their prior experiences [4].

Our study revealed that annotator groups generally used the tweet's link as supporting evidence for their arguments. This suggests that annotators analysed the tweet's context in addition to its face-value meaning at some point. Our findings also suggest that males are more likely than women to explore a link when looking for evidence to support a claim. White women and Black people were more prone to point to the use of racist and sexist terminology to justify their positions. This appears to contradict our hypothesis that Black women would rely more often on contextual features (such as; the tweet itself, the dialogue that a given tweet is a part of, other user bio details etc.). Our study also indicated that Black annotators have a higher tendency to draw from personal experiences to support their justifications. Another of our hypotheses was that Black women might rely more on personal experiences. Though the results from the participant's sources of justification suggest Black men are more likely to draw from their experiences than Black women, our examination and interpretation of the justifications (see Section 5.6) suggests otherwise, as Black women drew mostly from their personal experiences than groups other than Black women. However, this suggests the probability that participants may have indicated that their decision was supported by some type of evidence, which may not be reflected in their actual justification. To determine whether or not this is the case, it will be

necessary to investigate and analyse the participants' justifications in depth.

Therefore, a full qualitative analysis would be beneficial to be able to determine which interpretations were more prevalent among which groups and which sources of evidence appear in their justifications. As a result, future work will be directed toward conducting a comprehensive qualitative analysis of the 2,400 justifications and sources of evidence for the justification. In addition, we will further look at expanding the dataset and also investigating additional platforms.

This study has some limitations. The majority of the Black annotators are from South Africa, whereas the majority of the White annotators are from Europe, which may potentially skew the sample as there is the potential that their responses are a reflection of their cultural differences. Second, the results may potentially be skewed by the content diversity and the domain since we conducted an in-depth examination of 30 tweets from a single platform, Twitter, centred on four prominent Black women in technology whose experiences may be characterised as misogynoir. Some annotators found the online study (i.e. the annotation task) quite lengthy and often took more time than estimated, potentially impeding the annotator's ability to comprehend the work.

Despite these limitations and our open research directions, this study revealed that content comprehension and varied perspectives from diverse demographics in data annotations are relevant to understanding better intersectional hate (i.e Misogynoir) and the design of more effective detection approaches. We hope this analysis can incentivise the research community to investigate this phenomenon further.

7 CONCLUSION

In this paper, we explored the differences in how Black women annotate a dataset for potential cases of Misogynoir and Allyship compared with other groups. We analysed the variations in how our different participant groups annotate, the degree of confidence in their annotation, and the sources of evidence they employed in the annotation justification. In addition, we proposed a metric to detect contested tweets and to investigate further the tweets and the different interpretations derived from the annotator rationale to understand how meaningful these disagreements are. We sampled 30 tweets, designed an online annotation task study and recruited 80 annotators from the Prolific crowd-sourcing platform, 20 each from different ethnicity and genders including Black women, Black men, White women and White men.

We found that participants' sources of evidence in judging and interpreting content for potential cases of Misogynoir and Allyship, even in circumstances, where they all agree on a prospective label, vary across different factors, such as different ethnicity and gender. Given these identity factors, we presented how they influence the various interpretation of the content and sources of evidence for these interpretations.

This study demonstrates that further efforts are needed to understand better intersectional hate and the relevance of different annotator perspectives and content comprehension in hate speech.

8 ETHICAL CONSIDERATIONS

Considering the fact that our experiments expose annotators to potentially toxic and triggering content, we included a content warning and links to support groups on the survey instrument screen, describing the rating task and the possible harms that may result from participation. We asserted this at the beginning of the study:

> Risks associated with this study include feeling triggered or possibly harmed by viewing potentially toxic or hateful remarks and recalling unfavourable prior encounters with toxic online comments.

At the conclusion of the survey instrument, we supplied annotators with access to support pages where they may obtain assistance if they are triggered. We asserted this at the end;

> In the event that you are triggered by reading about these unpleasant experiences, we provide a list of resources to get support.

The names of the four Black women have been omitted to prevent identification and potential harassment. In accordance with Twitter's Terms of Service on Privacy and Anonymity, all shared and referenced tweets in this article have been rephrased.

ACKNOWLEDGMENTS

This work has received funding from the Melete Scholarship Scheme for innovation under the Melete Foundation Programme[9]. This work reflects only the authors' views, and Melete Foundation is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] Ishaan Arora, Julia Guo, Sarah Ita Levitan, Susan McGregor, and Julia Hirschberg. 2020. A Novel Methodology for Developing Automatic Harassment Classifiers for Twitter. Association for Computational Linguistics (ACL), 7–15. https://doi.org/10.18653/v1/2020.alw-1.2

[2] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. 25–35. https://doi.org/10.18653/v1/w19-3504

[3] Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling Bias in Toxic Speech Detection: A Survey. *Comput. Surveys* (1 2023). https://doi.org/10.1145/3580494

[4] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Proceedings of the 17th Symposium on Usable Privacy and Security, SOUPS 2021*. 299–317. https://data.esrg.stanford.edu/study/toxicity-perspectives

[5] Joseph Kwarteng, Serena Coppolino Perfumi, Tracie Farrell, and Miriam Fernandez. 2021. Misogynoir: Public online response towards self-reported misogynoir. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2021*. Association for Computing Machinery, Inc, 228–235. https://doi.org/10.1145/3487351.3488342

[6] Joseph Kwarteng, Serena Coppolino Perfumi, Tracie Farrell, Aisling Third, and Miriam Fernandez. 2022. Misogynoir: challenges in detecting intersectional hate. *Social Network Analysis and Mining* 12, 1 (12 2022), 1–15. https://doi.org/10.1007/s13278-022-00993-7

[7] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (3 1977), 159. https://doi.org/10.2307/2529310

[8] Marco Niemann, Jens Welsing, Dennis M. Riehle, Jens Brunk, Dennis Assenmacher, and Jörg Becker. 2020. Abusive Comments in Online Media and How to Fight Them. Springer, Cham, 122–137. https://doi.org/10.1007/978-3-030-61841-4{_}9

[9] https://melete.foundation/

[9] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *25th International World Wide Web Conference, WWW 2016*. 145–153. https://doi.org/10.1145/2872427.2883062

[10] Desmond U Patton, Philipp Blandfort, William R Frey, Michael B Gaskell, and Svebor Karaman. 2019. Annotating twitter data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, Vol. 2019-Janua. 2142–2151. https://doi.org/10.24251/hicss.2019.260

[11] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

[12] Ken Rogerson and Aidan Fitzsimmons. 2022. Intersectional Identities and Machine Learning: Illuminating Language Biases in Twitter Algorithms. In *Proceedings of the 55th Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2022.356

[13] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. (1 2017). https://doi.org/10.17185/duepublico/42132

[14] Gina Roussos and John F. Dovidio. 2018. Hate Speech Is in the Eye of the Beholder: The Influence of Racial Attitudes and Freedom of Speech Beliefs on Perceptions of Racially Motivated Threats of Violence. *Social Psychological and Personality Science* 9, 2 (1 2018), 176–185. https://doi.org/10.1177/1948550617748728

[15] Yisi Sang and Jeffrey Stanton. 2021. The Origin and Value of Disagreement Among Data Labelers: A Case Study of the Individual Difference in Hate Speech Annotation. (12 2021). https://doi.org/10.48550/arxiv.2112.04030

[16] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2020. The risk of racial bias in hate speech detection. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. Association for Computational Linguistics, 1668–1678. https://doi.org/10.18653/v1/p19-1163

[17] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. 5884–5906. https://doi.org/10.18653/v1/2022.naacl-main.431

[18] Trudy. 2014. Explanation Of Misogynoir. *Gradient Lair* (2014). http://www.gradientlair.com/post/84107309247/define-misogynoir-anti-black-misogyny-moya-bailey-coined

[19] Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science* 7 (6 2021), 1–38. https://doi.org/10.7717/PEERJ-CS.598