2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Annotators' Perspectives: Exploring the Influence of Identity on Interpreting Misogynoir

Joseph Kwarteng*, Tracie Farrell†, Aisling Third†, Miriam Fernandez†
Knowledge Media Institute, The Open University, UK
{joseph.kwarteng, tracie.farrell, aisling.third, miriam.fernandez}@open.ac.uk

***Abstract*—Social Networking Sites are home to different forms of hate, including "Misogynoir", which specifically targets Black women through a combination of racism and sexism. Detecting misogynoir presents challenges due to its subjective nature and the varied interpretations of hate speech. Using annotator justifications from four distinct demographic groups; including Black women, Black men, White women and White men, we seek to gain a deeper understanding of the factors that influence annotators' reasoning process and labelling decisions for potential cases of Misogynoir and Allyship. Given the unique experiences of Black women who face both racism and sexism, the study sought to understand how their intersectional identities shape their perspectives compared to other groups. The research employed a qualitative analysis of responses from participants to identify key themes and patterns. Three significant themes emerged from our in-depth qualitative analysis of these annotator justifications: prior knowledge and experience, the language of the social media post, and its context. Our results revealed that annotators historically at risk of abuse demonstrated a nuanced understanding of how their intersecting identities inform their interpretations and judgement of tweets, drawing on their personal encounters with misogyny and racism compared to their non-target counterparts of this type of hate. This study underscores the significance of diverse annotator perspectives and content comprehension in understanding and addressing hate speech, particularly when it intersects with multiple forms of discrimination. Our study contributes to the methodological advancements in social network analysis and mining, highlighting the importance of considering annotator characteristics in the development of tools and approaches for detecting and addressing intersectional hate.**

***Index Terms*—Misogynoir, Intersectionality, Social Media, Annotations, Hate Speech**

I. Introduction

In recent years, social networking platforms have become significant spaces for social interactions, influencing public discourse and societal conventions. This digital landscape has also become a breeding ground for various forms of hate, including "Misogynoir"; a term coined by Black feminist scholar Moya Bailey [1], [2] to describe the anti-Black racism and misogyny that Black women experience. As Black women's lived experiences shape their perspectives and contribute to a distinct understanding of social issues [3], their experiences of misogynoir further highlight the unique challenges they face in navigating digital platforms. Detecting misogynoir presents challenges as prior studies have found existing hate speech detection systems ineffective [4], [5] due to misogynoir's subjective and nuanced nature, and the varied interpretations of such hateful content across various demographics [6], [7]. As such, it is important to consider the characteristics of annotators to understand and interpret hateful speech [8]–[10] as these systems rely heavily on human-annotated datasets to learn from. Expanding on prior research [11] about different annotation behaviour of four distinct demographic groups: Black women (BW), Black men (BM), White Women (WW), and White men (WM), we aim with this current study to determine the driving influences of different annotator's distinct labelling judgements. Specifically, in this qualitative study, we seek to understand *RQ: How are Black women's reasoning for why a tweet could be misogynoir or allyship distinct from those of other groups?* We hypothesise that Black women may rely more on their lived experiences and knowledge of racist and sexist discourse. Studies like [12] have shown a link between annotators' identities and their perceptions of toxicity, hinting at potential biases of White annotators with prejudiced views. Similarly, [13] suggests that non-White annotators might be more attuned to subtle nuances in hate speech.

For this purpose, we used a previously collected dataset of 2400 rich annotator justifications from [11] about why annotators considered a post as misogynoir, allyship, unclear and none of the above. The content used to gather this data was from four identified misogynoir cases in the Tech sector involving four prominent Black women collected from Twitter (§ III-A). Our qualitative analysis of annotator justifications revealed three key themes: prior knowledge and experience, the tweet's language (linguistic aspects of the content) and contextual information. Notably, annotators who are historically at risk of abuse demonstrated a nuanced understanding of how their intersecting identities shape their interpretations of tweets, drawing on personal and lived experiences with misogyny and racism. In contrast, non-target annotators lacked this level of depth in their assessments. These findings underscore the significance of personal experiences in influencing annotators' comprehension of content.

This study received funding from The Melete Foundation under the Melete Scholarship Scheme for Innovation Programme.

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11…$15.00
https://doi.org/10.1145/3625007.3627292

Our contributions can be summarised as follows:

- A novel approach to understanding annotator influence in the context of misogynoir centred on the justification by four distinct demographic groups.
- A rich qualitative data of 2400 annotator justifications generated by 80 annotators from the four distinct demographic groups.
- A comprehensive qualitative analysis of the specific factors participants named as influencing their reasoning and labelling of content as misogynoir and allyship, contributing to the advancement of knowledge around intersectionality in online hate.

The rest of the paper is structured as follows. (§II) describes relevant related work. (§III) describes our analysis setup for the study. Results of the analysis are presented in (§IV). Discussions and conclusions are presented in (§V) and (§VI), respectively[1].

II. Related Work

Existing studies have explored the intersectionality of race and gender, particularly concerning the experiences of Black women [3], [14]–[16]. Crenshaw, in her introduction to the concept of intersectionality, argued that Black women's experiences cannot be understood solely through the lens of either race or gender, but rather through the intersection of these identities [3], [15]. This raises the question of whether those who never experienced misogynoir can comprehend its complete impact and significance.

In combating hate speech, annotation collection is crucial for developing effective tools. However, it poses challenges as people's perceptions of hateful content differ based on their demographics and prior experiences [6]. Prior studies found factors such as abuse victimhood, ethnicity, racial beliefs, context, lived experiences, and gender to have a significant impact on how people perceive and interpret hate speech and how they label it [8]–[11]. Research indicates victims and targets who have directly experienced abuse in the past are more likely to label a statement as toxic, as well as groups historically at risk of abuse [8]. Annotators' identities and beliefs, as well as annotators' ratings of toxicity, are strongly correlated, as White annotators, particularly those with racist beliefs, may be unreliable in annotating racist toxicity online [12] and non-White annotators may be more sensitive to nuances in hate speech [13]. However, studies have shown limited focus on annotator positionality, which refers to how an annotator's social identity influences their perception and understanding of the world [17], [18].

Likewise, the concept of allyship with Black women has been a subject of scholarly and feminist debate, with variations in terminology. Allies are typically individuals from privileged groups who leverage their majority status to promote positive change [19], [20]. Allyship for Black women would involve privileged individuals supporting and advocating for them to address their specific challenges and injustices [21]. However, many critiques exist from Black women (and more generally, People of Colour) around the performative or surface nature of allyship from White People (and White Women specifically) [22]. We, therefore, examined justifications of Allyship messages to explore how the different groups view this issue.

Building on this existing scholarship, this study aims to examine the distinct reasoning of Black women regarding misogynoir and allyship compared to other demographic groups. Some of these factors elaborated in prior research may play a role in what informs the annotators' decisions in labelling a tweet as a potential instance of misogynoir. By comparing the responses of Black women, Black men, White women, and White men, this study seeks to dig deeper into what factors influence the reasoning process.

III. Analysis Setup

In III-A, we provide an overview of the dataset, explain our research approach in III-B, and outline the data analysis methods in III-C. We also describe the themes and codes generated from the analysis in III-D.

A. Data

In this study, we utilised a dataset of 2400 annotator justifications, exploring reasons for classifying tweets as either Misogynoir or Allyship, from a prior study [11]. These justifications came from 80 annotators, including 20 individuals each from four demographic groups: BW, BM, WW, and WM, recruited via Prolific [2]. The original study by [11] employed a web survey to gather participants' rationales when annotating potential misogynoir tweets. Subsequently, participants classified and justified their categorisation of 30 tweets into four categories: Misogynoir (M), Allyship (A), Unclear (U), or None of the Above (NA). In this paper, we centred our attention on the reasons (justifications) annotators provided for categorising tweets as either Misogynoir or Allyship.

B. Research Approach

This study employed a qualitative research method to explore the factors that influence participants' labelling decisions and interpretations of misogynoir and allyship. Motivated by the desire to identify and interpret patterns, themes, and meanings within the qualitative data, thematic analysis [23] was employed as the primary methodological approach. As a systematic yet flexible approach to analysing textual or verbal data, thematic analysis enables the investigation of underlying themes and captures the nuances and complexity of participants' perspectives [24]. In addition, an inductive analysis approach was adopted to generate insights and theories directly from the data [25]. This facilitates a more open and exploratory analysis, which allows emergent themes and patterns to be identified without being constrained by prior assumptions. Integrating thematic analysis with inductive analysis ensures a rigorous and thorough examination of the qualitative data, facilitating a deeper understanding of the

[1]The code and the data used for the analysis will be made publicly available under https://github.com/kwartengj/Asonam23_Misogynoir

[2]https://www.prolific.co/

differences in influences of these different participant labelling decisions.

C. Data Analysis

In this study, we followed the six-step framework approach for conducting a thematic analysis by [24]. The initial phase of data analysis consisted of a comprehensive reading of the data to become familiar with the data and generate initial ideas and thoughts. These initial concepts or notions are then put into well-defined and demarcated codes. Line-by-line coding was then conducted, where meaningful segments of the data were assigned descriptive codes (§ III-D). This process involved identifying recurring patterns, concepts, and ideas within the dataset. In the context of meaningful segments; we set our unit of analysis as "complete thought". Thus, where the participant provided a lengthy justification for their decision, there might be more than one complete thought expressed within that justification. The full justification might have more than one code attached to it but each complete thought has one code attached to it. Once the initial coding was completed, codes were grouped according to the degree of similarity between them, resulting in the development of preliminary themes. These preliminary themes were further refined through a process of comparison, discussion, and revisiting of the data. This iterative approach allowed for the identification of overarching themes that captured the essence of the participants' experiences and perspectives (§ III-D for themes and code descriptions).

To increase the credibility and reliability of the findings, a team-based approach was utilised. Multiple researchers (authors of this paper) were involved in the generation of the codebook through iterative cycles of coding and review. Regular team meetings were conducted to discuss and review emergent themes and new codes, and to resolve disputes about the codes, their definitions, or the comparative examples we used to help code further data. Throughout the data analysis process, an audit trail was maintained to ensure transparency and rigour. Detailed documentation of coding decisions, coding memos, and reflective notes was captured to provide a transparent record of the analysis process.

D. Themes and Codes

1) Prior Knowledge and Experience: This refers to the existing knowledge and experiences that individuals bring to their interpretation of the tweets. The analysis delves into how this prior knowledge and experience shapes individuals' recognition and interpretation of misogynoir, as well as their ability to identify and challenge misogynoiristic language and attitudes. See Table I for generated codes under this theme.

2) Tweet's Language: This theme refers to the linguistic characteristics and elements present in the analysed tweets. The analysis examines the impact of derogatory language, stereotypes, slurs, and other linguistic elements that contribute to the expression of misogynoir in the tweets on participants' understanding, perception, and labelling decisions. See Table II for generated codes under this theme.

TABLE I
CODES UNDER THE PRIOR KNOWLEDGE AND EXPERIENCE THEME

Codes	Definitions
Lived experience	This code is for statements where the participants directly reference their own lived experiences in the content as grounds for their interpretation of their justifications.
Knowledge about racist and sexist discourse	This code is for statements where the participants do not talk about personal experiences but express some prior knowledge about racist and sexist discourse.
Participants' opinions	This code is for statements where the participants did not name any specific type of evidence but expressed a personal opinion or intuition about the tweet.

TABLE II
CODES UNDER THE TWEET'S LANGUAGE THEME

Codes	Definitions
Presence of racist or sexist comments	This code is for statements where the participants state that their justification is based on the presence of explicitly racist or sexist terms.
Absence of racist or sexist comments	This code is for statements where the participants directly reference the absence of specific racist or sexist terms as the grounds for their interpretation of their justifications.
Perceived attack on a Black woman	This code is for statements where the participants perceive an attack on a person on the basis that the person is a Black woman (If the justification is attached to an annotation of misogynoir, it can be presumed that the person feels this attack is on the basis of the person's being a Black woman, whether they explicitly state this or not).
Perceived absence of an attack on a Black woman	This code is for statements where the participant explicitly states that they perceive no attacks on a person on the basis that the person is a Black woman.
Specified language	This code is for statements where the participants reference or mention specific terms or phrases but do not identify them as specifically racist or sexist.
Support Language	This code applies to statements for which the participants' justification is that the statement demonstrates solidarity or support for the victim.
Unspecified language	This code is for statements where the participants reference something about the tweet's language but are not that clear or specified as to what it is. (This will include all justifications around tone that are not accompanied by evidence of that tone - such as would be provided in the following code).

TABLE III
CODES UNDER THE CONTEXT THEME

Codes	Definitions
Deleted text or content	This code is for statements where the participants directly reference the deletion of a tweet as the grounds for the decision.
Lack of clarity	This code is for statements where the participants based their decision on the tweet being unclear, not understandable or confusing.
Author characterisation	This code is for statements where the participants directly reference historical data about the tweet's author (previous tweets, information in bio, etc.), or their assumptions about them.
Unspecified author characterisation	This code is for statements where the participant mentions something about the author (assumptions about the author and what they said) without any given evidence.
Needs further information	This code is for statements where the participants mention the need for more details or extra information.

3) Context: This theme delves into the analysis and comprehension of contextual factors that aid in the interpretation of tweets, as well as their impact on participants' labelling judgments. It considers the broader concept of what was said, who said it, and the historical data (i.e. previous tweets, information on his/her bio etc) of the person who said it, as well as what part of the context is missing that hinders the reader's comprehension. See Table III for generated codes.

IV. Analysis Results

In this section, we present the results of our qualitative analysis, guided by our research question; ***RQ: How is Black women's reasoning for why a tweet could be misogynoir or allyship distinct from those of other groups?***

A. Misogynoir

1) Prior knowledge and Experience: When participants were asked about their experiences with misogynoir (see Table IV), we observed a higher number of participants from Black women (12 participants) and Black men (16 participants) reporting personally experiencing and witnessing misogynoir respectively, in comparison to White women (2 participants) and White men (9 participants).

TABLE IV
Participants' responses about their experiences with Misogynoir

Have you..	Number of Participants			
	BW	***BM***	***WW***	***WM***
Witnessed misogynoir	5	16	8	9
Personally experienced misogynoir	12	-	2	1
Not sure	2	4	2	5
None of the above	1	-	8	5

Relevant examples provided by the two White women who claimed to have personally experienced misogynoir are more consistent with misogyny than with misogynoir, which is exclusive to Black women according to its definition [1]. E.g. one of the White women stated, *"My ex-husband didn't let me work. He wanted to make money for us. After our quarrel, he reminded me that I am nobody because I do not earn"*. Likewise, the White man who asserted personal experience with misogynoir stated, *"in the sexual act"*; which is difficult to interpret because White men per definition do not experience either misogyny or misogynoir.

In terms of coded frequencies, Black women had the highest occurrence with 60 instances, followed by Black men with 45, White women with 29 and White men with 15. This suggests that Black women and men may have relied on a broader range of prior knowledge and experiences compared to their White counterparts.

The frequencies of **knowledge about racist and sexist discourse** varied among the different demographic groups, with Black women having the highest occurrence, followed by Black men, White women, and White men with decreasing frequencies (see Fig. 1). Black women's justifications include discourses about microaggressions, the invalidation of their feelings through sarcasm and dismissive phrases and the angry

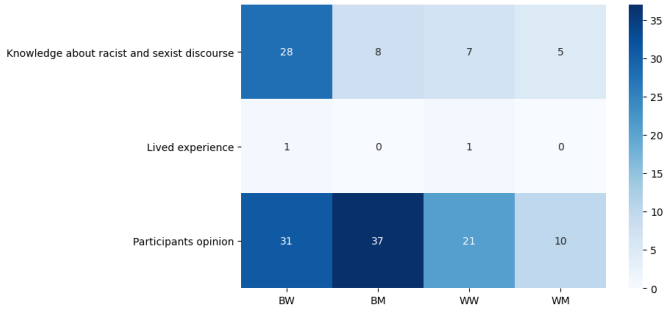

Fig. 1. Number of Coded References in the Prior Knowledge and Experience Theme for Misogynoir

Black woman narrative (e.g., one Black woman wrote *"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"*, whereas Black men's justifications include discourses about undermining remarks and silencing tactics (e.g., one Black man wrote *"Black women are always silenced by being constantly told that they are toxic when they demand the same respect others get"* and White women's justifications include discourses about sexism and using derogatory remarks to invalidate Black women's feelings (e.g., one White woman wrote *"... Every time black women speak up about their experience with racism and sexism, they get shot down and labelled aggressive and emotional."*) and the justification of White men centres around sexism (e.g., one White man wrote *"The Twitter user, uses the argument that she was lucky to get an opportunity so she shouldn't demand anything else. It is an argument fairly used against women"*). All of the above suggests that Black women may possess a deeper understanding of the nuances of racist and sexist discourse around misogynoir than the other groups.

Under **lived experiences**; we only had Black women and White women reporting one instance each (see Fig. 1). This may be due to the fact that under this code, we were looking for explicit references to how the participants' experiences informed their justifications. The Black woman's justification reads *"... 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"* while the White woman's justification reads *"This tweet suggests that the woman has a bias complex about her skin color. In my own experience, I believe that black women experience persecution, so it's not a complex issue"*. The Black woman's justification reflects the influence of personal experiences and the understanding that, as a Black woman, she faces challenges in having her perspective acknowledged, whereas the White woman's perspective may stem from her understanding of broader racial dynamics and biases.

Under **participants opinion**; Black men reported the highest occurrence, followed by Black women, White women and

White men with progressively lower frequencies (see Fig. 1). This suggests that Black participants may place a stronger emphasis on their opinions or intuitions compared to the other groups. Black women's justifications ranged from unfair treatment, invalidating the victim's feelings and experiences, and certain remarks about the victim's race and the race of the author of the tweet (e.g., one Black woman stated *"..., the author being of a different race, I doubt that they would see anything wrong about the unfair dismissal"*), whereas Black men's justifications included: lack of supportive remarks, criticising without understanding (e.g., one Black man stated *"She is commenting without really understanding the situation, she is merely saying this since the black woman is complaining about unfair treatment"*), silencing or oppressing Black women, and making a joke about the situation. White women's justification included; impolite writing styles (e.g., *'... The person even uses caps, trying to emphasize "his point"*), identified sexist intents and the lack of support, while White men's justification also included; attempts to downplay the problem and unfair treatment (e.g., one White man said *"...the tweet was really unfair and inconsiderate in regards to the person he was talking about."*) Each group's justifications seem related to their own experiences of gender- and/or race-based discrimination, e.g., Black women by misogynoir, Black men and White women by race and gender, respectively. White men by their limited understanding of misogynoir as non-targets or as potential perpetrators.

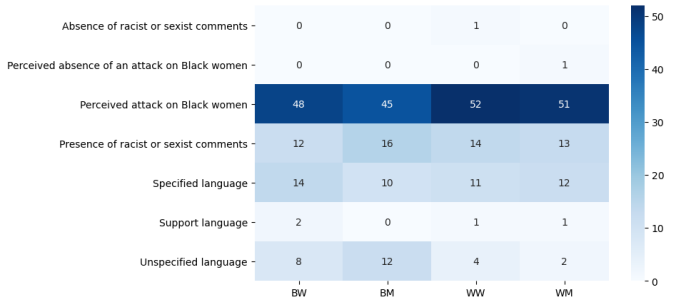

Fig. 2. Number of Coded References in the Tweet's Language Theme for Misogynoir

2) Tweets' Language: The sum of coded frequencies across the different groups under this theme is generally similar ranging from White men 80 instances to Black women 84 and both White women and Black men having 83 each. This suggests a shared recognition of the importance of analysing the language related to misogynoir, which is expected given that we asked that they analyse tweets.

Under the **Absence of racist or sexist comments** and the **Perceived absence of an attack on Black women**, only a White woman and a White man reported a single occurrence. The White woman stated *"Although the user doesn't openly use any words or phrases that would indicate their racism or sexism"* and the White man also stated *"I don't see anything misogynoir or harmful with the mentioned tweet."*

Under the **Perceived attack on Black women** code; the frequencies among the groups are relatively similar, with slight variations, ranging from 45 to 52 instances (see Fig. 2) with White women having the highest (52) followed by White men (51), Black women (48) and Black men (45). This suggests a comparable understanding of attacks on Black women across different demographic groups. Black women's justifications included attacks on the victim and their race, as well as displays of disrespect and hostility (e.g., one Black woman wrote *"The statement is very harsh towards the fired person and passing judgment to a person this author has never met. It makes me feel like the fired person was not supposed to question anything because she is a woman of colour."*, whereas Black men's justifications included prejudice against the victim, victim-blaming, and not believing Black women (e.g., one Black man wrote *"This shows hatred he has against her, the person who made the tweet did not support or believe that the girl was oppressed, he just concluded that she wrote the tweets about oppression because she was looking for attention, and this is because she was black"*. White women's justifications included discourse about prejudice against the victim (e.g., attacking their race and skin colour) and mocking Black women's experiences (e.g., one White woman wrote *"The author of the tweet shows clear personal racist bias. The argument that Black people do crime because they're Black, and not because they're victims of the racist imbued systems, is a pretty old argument"*, while White men's justifications included discourse about prejudice against the victim (such as having ill intent and attacking race with aggression and anger) (e.g., one White man wrote *"...person who wrote that tweet had ill-intentions and was really dismissive about what @user might have been feeling"*. These differences in justifications indicate the influence of intersecting identities and societal dynamics on individuals' interpretations and responses to attacks on Black women.

Regarding the **Presence of racist or sexist comments** code, all groups recognised tweets containing such comments, with slight variations in frequencies. Black men reported 16 instances, followed by 14 from White women, 13 from White men, and 12 from Black women (see Fig. 2). E.g., one Black woman stated, *"The comment has some sexist and racist features"*, while a Black man wrote *"It shows racist comments which a horrible"*, a White woman stated *"I consider it an explicit racist comment"* and a White man also noted *"Racist phrases in the tweet make me sure that this tweet is misogynoir"*. These findings demonstrate a collective awareness of discriminatory language in misogynoiristic tweets.

Under the **Specified language** and the **Unspecified language**; the frequencies of specified language varied among the groups, ranging from 10 to 14 with Black women having the highest (see Fig. 2). This suggests that participants from various demographics paid attention to specified language in the tweets. The frequency of unspecified language varied across the groups: Black men had the highest occurrence (12 instances), followed by Black women (8), and White women and White men had (4) and (2) respectively. Black participants and White women's justifications included discourses around gaslighting language and how the comment is phrased (i.e. the

tone and the structure and writing styles of the comment) e.g., one Black woman wrote *"Another case of gaslighting victims and being rude towards them"*, one Black man wrote *"The tone in the text is very sexist, even the tweeter handle tells you that this person is racist and they do not like black women"* and one White woman also wrote *"This sentence has negative overtone"*. White men's justifications were more about intent or feeling e.g., one White man wrote *"I feel some malicious edge in this tweet"*. Black women, Black men, and White women experience prejudice (racism and misogyny) relative to White men, which could be an explanation for these findings. Moreover, results from unspecified language suggest that misogynoir is nuanced. People may be catching up on a combination of terms or the use of particular language conventions rather than specific hateful terms.

3) Context: This theme includes two sub-themes: *"Author Specification"* and *"Knowledge of Society"*. "Author specification" refers to explicit and non-explicit details about the tweet's author, such as language, tone, and references to historical data. "Knowledge of society" relates to individuals' understanding of social dynamics, power structures, inequalities, systemic inequalities and cultural norms. Among the groups, Black women had the highest frequency of occurrences (27 instances), followed by White women (21), Black men (18), and White men (9) in the context theme (see Fig. 3). This suggests that Black women emphasize context more frequently in their justifications compared to other groups, while White men stand out as distinct.

Author Specification appears to be present across all four groups, with varying frequencies. In terms of **"Author characterisation"**, White women had the highest occurrence of this code with 9 instances, followed by Black women with 7, White men with 5, and Black men with 4. With **"Unspecified author characterisation"** Black women had the highest frequencies with 15 instances, followed by White women with 13, Black men with 11 and White men with 2. Justifications provided under this theme had comparable rationales from all demographic groups; as to either checking the author's previous tweets, and bio or combining the author's specification with their knowledge of society. For example one Black woman wrote *"I label this as misogynoir because of @user's other tweets. He seems to have an arrogance and an unforgiving nature when it comes to other peoples issues..."*, one Black man also wrote *"The author of the tweet is a racist who believes minorities must not stand up for themselves"*, one White woman also wrote *"Based on the tweet handle, it seems quite obvious the person who tweeted was sexist..."* and one White man wrote *"The tweet was already pretty misogynoir sounding but looking at their profile there were some horrible and straight up racist stuff"*. These results suggest that women are nearly twice as likely to utilise context-related information to comprehend tweets and classify them, even though all groups to some extent weighed the author's character and knowledge of society.

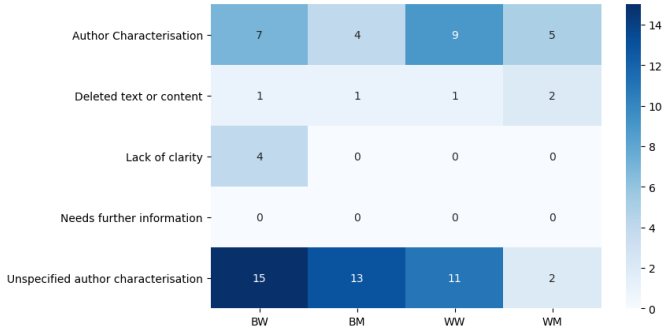

Fig. 3. Number of Coded References in the Context Theme for Misogynoir

B. Allyship

While expressions of hate can vary significantly depending on the target, allyship tends to show more consistency. In our dataset, participants displayed a shared understanding and experience of allyship, particularly in terms of expressing gratitude and sharing personal experiences which is easy to spot linguistically. This could be that expressions of allyship may not explicitly reference the specific features of attack/hate speech - perhaps allyship may be expressed in similar ways across various forms of prejudice that people have seen more examples of it in more context, and beyond race and gender.

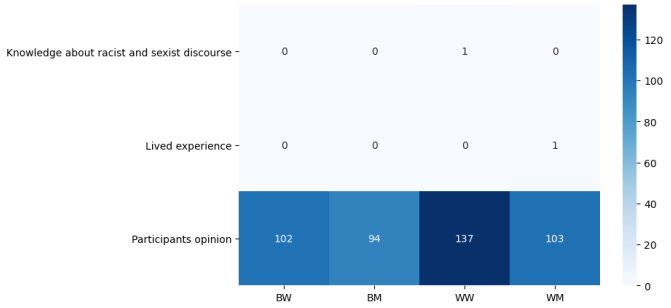

Fig. 4. Number of Coded References in the Prior Knowledge and Experience Theme for Allyship

1) Prior knowledge and Experience: The results for this theme appear to be a slight mirroring of what we see in misogynoir but perhaps there is less attention to detail here as long as the tweet is perceived to show support or gratitude or defend the victim (Black women). In summary, all groups primarily rely on their own opinions, intuitions, and perspectives when evaluating the allyship qualities of a tweet, with little mention of knowledge about racist and sexist discourse or lived experiences (see Fig. 4). The justifications provided had similar opinions among all demographic groups, including tweets expressing concern, positivity, support, and encouragement. Also, authors share similar experiences, thanking the victim for sharing theirs, the use of hashtags, and standing in solidarity with the victim. One Black woman wrote *"They seem to be expressing concern with @user and they want to connect with her. The tweet comes across as of concern."*, one Black man wrote *"The hashtags. The overall tweet and*

message". One White woman also wrote *"Very clearly positive and encouraging tweet,... Also the expression of personal gratitude."* and a White man wrote *"Thanking her for sharing the information."*

2) Tweets' Language: In summary, the findings indicate that Black participants placed a greater emphasis on language-related factors associated with support; thus language expressing support, empathy, and understanding as a crucial element of allyship. **"Support language"** is the predominant code across the theme with frequencies higher among Black women (141 instances) and Black men (156) as compared to White women (94) and White men (102) (see Fig. 5). One Black woman wrote *"Very clear allyship. this person is being supportive..."*, and one Black man wrote *"Offers support and advice"*. One White woman also wrote *"Shows support"* and one White man wrote *"Again a show of sympathy."*

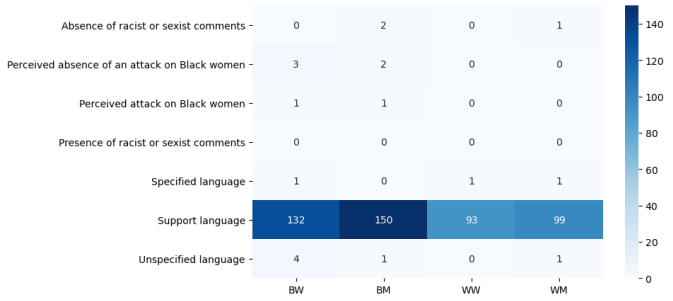

Fig. 5. Number of Coded References in the Tweet's Language Theme for Allyship

3) Context: The results for this theme appear to be a slight reflection of what we observe in misogynoir, but perhaps there is less focus on the authors' historical data and more on what participants viewed the authors of the tweets to be and their interpretation of what is being said (i.e., are they perceived to be an ally, and acting as expected for an ally). Notably, women (Black: 43 and White: 49) had higher frequencies compared to me (Black: 23 and White: 12). This suggests that women may engage more in making generalisations or assumptions about the tweet author's characteristics when evaluating allyship compared to men. The justifications provided by the different demographic groups included defending the victim (Black women), offering support or sympathy, using supportive hashtags, and showing solidarity. One Black woman wrote *"The author suggests that the is a lot wrong and unjust in the world and that people try to hide the truth"*, and one Black man wrote *"This is an ally, and they also clearly experience the same issues"*. One White woman also wrote *"In addition to the comment, the nature of an ally becomes even clearer through the #"* and one White man wrote *"shares his observations of the inefficiency in the fight against racism... he is an ally."*

V. Discussion

Previous studies have established the importance of annotator characteristics, such as examples of ethnicity and gender, in the evaluation of hate speech [8]–[10]. Moreover, research has shown that annotators from different racial and gender

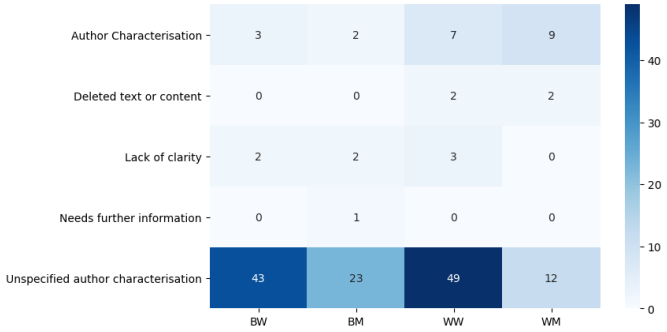

Fig. 6. Number of Coded References in the Context Theme for Allyship

backgrounds, even when in agreement on a label, consider distinct types of evidence [11]. Building upon this existing literature, our study goes deeper into the actual justifications that inform annotators' labelling decisions to really pinpoint how lived experience makes the difference. Our study delved deeper into the factors that influence annotator labelling decisions, specifically emphasising the significance of prior knowledge and experience, linguistic aspects of the content, and contextual information.

Compared to the other demographic groups, we found that Black women's justifications were more representative across the three key elements. Their frequent and detailed justifications provided strong evidence of their profound understanding of the subject matter, which stemmed from their lived experiences of misogynoir [3], [14]–[16]. The lived experience encompasses a deeper knowledge of the discourse surrounding misogynoir, heightened awareness of the use of specific terms and phrases, and a greater understanding of the contextual dynamics of hate and the identities of its perpetrators. Conversely, participants from other groups exhibited varying levels of knowledge and provided justifications that were comparatively less rooted in personal experiences, which support and validate our initial hypotheses. This highlights the inclusion of Black women in discussions on misogynoir as crucial, given their unique experiences and deep understanding of the subject matter [18].

Interestingly, our analysis also indicated some similarities in reasoning between Black women and other groups. For example, both Black women and non-Black women recognised the significance of harmful stereotypes and attacks on Black women and the need for allyship. However, the nuanced understanding of intersectionality and the incorporation of historical context were distinctive elements within the reasoning of Black women. Among White groups, this could be attributed to their familiarity as historical perpetrators of this animosity, in contrast to Black groups, who are historical targets who experience this daily. Also, our study uncovered a notable gender difference, indicating that women are more inclined to consider contextual information, such as explicit and unspecified details of the tweet's author, compared to men.

The study also revealed some of the challenges of nuanced language when addressing intersectional hate, such

as misogynoir. The complex nature of intersecting forms of discrimination makes it difficult to capture and label instances of hate speech accurately [4], [11]. Annotators faced the challenge of interpreting nuanced language that encompassed both racism and sexism, highlighting the complexity of their identities [17]. This raises concerns about the effectiveness of the annotation process in addressing intersectional hate.

However, it is essential to acknowledge some limitations. The sample size was relatively small and may not be fully representative of the diverse experiences of Black women, especially as a significant portion of our Black annotators were based in South Africa, while the majority of our White annotators were from Europe. This geographical distinction could potentially skew perceptions and experiences. Also, the study focused specifically on Twitter, limiting the generalisability of the findings to other social media platforms. Future research should aim to include a larger and more diverse sample and explore different online platforms.

Despite these limitations, this study emphasises the influence of personal perspectives, cultural backgrounds, and social identities on individuals' interpretations of misogynoir. Our findings demonstrate that Black women possess a deep understanding of nuanced language and extensive knowledge of the historical and contemporary discourse on misogynoir. This underscores the significance of considering the unique perspectives and experiences of targets, (in this case, Black women) in discussions of intersectional discrimination.

VI. Conclusion

In this study, we sought to investigate how Black women's reasoning for categorising tweets as misogynoir or allyship differs from that of other groups. We used a dataset of 2400 annotator justifications from four distinctive demographic groups including; Black women, Black men, White women and White men. Through a thorough thematic analysis of the data, we identified three key themes that shed light on the unique perspectives and experiences of Black women in this context: prior knowledge and experience, linguistic aspects of the content, and contextual information. This study contributes to the existing discussion by demonstrating that annotators who are targets of abuse, particularly Black women, possess a nuanced understanding of how their intersecting identities influence their interpretations of tweets, drawing from lived experiences and knowledge of misogyny and racism. In contrast, annotators without target experiences lacked depth in their assessments. These findings emphasise the importance of lived experiences in shaping annotators' comprehension of content and their ability to identify instances of misogynoir and intersectional hate. Also, it highlights the significance of centring marginalised voices and lived experiences in the identification and addressing of intersectional hate.

VII. Ethical Considerations

Names or handles in the quoted justifications have been anonymised to "@user." Given the sensitive topic of Misogynoir, resources were provided for participants who might find the study distressing.

References

[1] Trudy, "Explanation Of Misogynoir," *Gradient Lair*, 2014. [Online]. Available: http://www.gradientlair.com/post/84107309247/define-misogynoir-anti-black-misogyny-moya-bailey-coined

[2] M. Bailey and Trudy, "On misogynoir: Citation, erasure, and plagiarism," *Feminist Media Studies*, vol. 18, no. 4, pp. 762–768, 2018.

[3] K. Crenshaw, "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics," in *Feminist Legal Theory: Readings in Law and Gender*, 1989, vol. 1989, no. 1, pp. 57–80. [Online]. Available: http://chicagounbound.uchicago.edu/uclfhttp://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

[4] J. Kwarteng, S. C. Perfumi, T. Farrell, A. Third, and M. Fernandez, "Misogynoir: challenges in detecting intersectional hate," *Social Network Analysis and Mining*, vol. 12, no. 1, pp. 1–15, 12 2022. [Online]. Available: https://link.springer.com/article/10.1007/s13278-022-00993-7

[5] K. Rogerson and A. Fitzsimmons, "Intersectional Identities and Machine Learning: Illuminating Language Biases in Twitter Algorithms," in *Proceedings of the 55th Hawaii International Conference on System Sciences*, 2022. [Online]. Available: https://hdl.handle.net/10125/79690

[6] T. Garg, S. Masud, T. Suresh, and T. Chakraborty, "Handling Bias in Toxic Speech Detection: A Survey," *ACM Computing Surveys*, 1 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3580494

[7] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 1668–1678.

[8] D. Kumar, P. G. Kelley, S. Consolvo, J. Mason, E. Bursztein, Z. Durumeric, K. Thomas, and M. Bailey, "Designing toxic content classification for a diversity of perspectives," in *Proceedings of the 17th Symposium on Usable Privacy and Security, SOUPS 2021*, 2021, pp. 299–317. [Online]. Available: https://data.esrg.stanford.edu/study/toxicity-perspectives

[9] Y. Sang and J. Stanton, "The Origin and Value of Disagreement Among Data Labelers: A Case Study of the Individual Difference in Hate Speech Annotation," 12 2021. [Online]. Available: https://arxiv-org.libezproxy.open.ac.uk/abs/2112.04030v1

[10] D. U. Patton, P. Blandfort, W. R. Frey, M. B. Gaskell, and S. Karaman, "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, 2019, pp. 2142–2151. [Online]. Available: https://hdl.handle.net/10125/59653

[11] J. Kwarteng, G. Burel, A. Third, T. Farrell, and M. Fernandez, "Understanding misogynoir: A study of annotators' perspectives," in *Proceedings of the 15th ACM Web Science Conference 2023*, 2023, pp. 271–282.

[12] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. Smith, "Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection," 2022, pp. 5884–5906.

[13] S. Larimore, I. Kennedy, B. Haskett, and A. Arseniev-Koehler, "Reconsidering annotator disagreement about racist language: Noise or signal?" in *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, 2021, pp. 81–90.

[14] P. H. Collins, "Intersectionality's Definitional Dilemmas," pp. 1–20, 8 2015. [Online]. Available: www.annualreviews.org

[15] K. Crenshaw, "Mapping the margins: Intersectionality, identity politics, and violence against women of color," *Stan. L. Rev.*, vol. 43, p. 1241, 1990.

[16] S. Bernstein, "The metaphysics of intersectionality," *Philosophical Studies*, vol. 177, no. 2, pp. 321–335, 2 2020. [Online]. Available: https://doi.org/10.1007/s11098-019-01394-x

[17] M. K. Scheuerman, A. Hanna, and E. Denton, "Do datasets have politics? disciplinary values in computer vision dataset development," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–37, 2021.

[18] R. S. Geiger, K. Yu, Y. Yang, M. Dai, J. Qiu, R. Tang, and J. Huang, "Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from?" in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 325–336.

[19] I. E. Sabat, L. R. Martinez, and J. L. Wessel, "Neo-activism: Engaging allies in modern workplace discrimination reduction," *Industrial and Organizational Psychology*, vol. 6, no. 4, pp. 480–485, 2013.

[20] N. P. Salter and L. Migliaccio, "Allyship as a diversity and inclusion tool in the workplace," *Diversity within Diversity Management*, vol. 22, pp. 131–152, 2019.

[21] M. M. Karnaze, R. M. Rajagopalan, L. T. Eyler, and C. S. Bloss, "Compassion as a tool for allyship and anti-racism," *Frontiers in Psychology*, vol. 14, p. 1143384, 2023.

[22] K.-Y. Taylor, *How we get free: Black feminism and the Combahee River Collective*. Haymarket Books, 2017.

[23] V. Clarke, V. Braun, and N. Hayfield, "Thematic analysis," *Qualitative psychology: A practical guide to research methods*, vol. 3, pp. 222–248, 2015.

[24] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[25] L. Varpio, E. Paradis, S. Uijtdehaage, and M. Young, "The distinctions between theory, theoretical framework, and conceptual framework," *Academic Medicine*, vol. 95, no. 7, pp. 989–994, 2020.