Misogynoir: Public Online Response Towards Self-Reported Misogynoir

Joseph Kwarteng*, Serena Coppolino Perfumi†, Tracie Farrell‡, Miriam Fernandez‡
* ‡ The Open University, United Kingdom
†Stockholm University, Sweden
{joseph.kwarteng, tracie.farrell, miriam.fernandez}@open.ac.uk
serena.perfumi@sociology.su.se

***Abstract*—"Misogynoir" refers to the specific forms of misogyny that Black women experience, which couple racism and sexism together. To better understand the online manifestations of this type of hate, and to propose methods that can automatically identify it, in this paper, we conduct a study on 4 cases of Black women in Tech reporting experiences of misogynoir on the Twitter platform. We follow the reactions to these cases (both supportive and non-supportive responses), and categorise them within a model of misogynoir that highlights experiences of Tone Policing, White Centring, Racial Gaslighting and Defensiveness. As an intersectional form of abusive or hateful speech, we investigate the possibilities and challenges to detect online instances of misogynoir in an automated way. We then conduct a closer qualitative analysis on messages of support and non-support to look at some of these categories in more detail. The purpose of this investigation is to understand responses to misogynoir online, including doubling down on misogynoir, engaging in performative allyship, and showing solidarity with Black women in tech.**

***Index Terms*—Misogynoir, Social Media, Public Response**

I. Introduction

The portmanteau "misogynoir" was coined in 2008 by Moya Bailey to describe the specific forms of misogyny that Black women experience in visual and digital culture, which are coupled with racism, as well as heterosexual desire and normative expressions of gender [1]. The term was further developed by Trudy (aka @thetrudz) [2][1] and the Crunk Feminist Collective[2] to include social or institutional environments [2], [1]. For example, hypersexualisation of Black women and stereotypes that characterise Black women, particularly, as angry, unreasonable, or unintelligent are examples of misogynoir that impact the health, safety and well-being of Black women and girls [3]. These biases are also visibly encoded in language [4]. Understanding misogynoir as a specific type of harm experienced by Black women is important for reshaping industries and fields with low representation.

[1]http://www.thetrudz.com/

[2]https://www.crunkfeministcollective.com/

ASONAM '21, November 8–11, 2021, Virtual Event, Netherlands

ACM ISBN 978-1-4503-9128-3/21/11/...$15.00
http://dx.doi.org/10.1145/3487351.3488342

Studies focused on the investigation of misogynoir in online environments (particularly social networks), provide in-depth observations of the rhetoric around misogynoir, but they are generally conducted manually and over small data samples [5]. In this work we aim to: (i) understand and characterise how misogynoir is manifested online and, (ii) investigate potential tools and methods to enable an automatic identification of this type of hate. Our contributions can be summarised as:

- An extensive study of the literature of misogynoir and the identification of five distinct themes (White Centring, Tone Policing, Racial Gaslighting, Defensiveness and General Commentary) from the social science and humanities literature.
- The translation of the different identified themes, into a lexicon of terms and expressions, to experiment with the automated identification and study of the different manifestations of misogynoir online.
- An in-depth quantitative and qualitative analysis of four high-profile cases of misogynoir in Tech (see Section IV-A), and a manually annotated dataset of 2,519 Twitter posts capturing public responses of misogynoir online (both supportive and non-supportive messages).
- An analysis of the challenges and opportunities for understanding misogynoir online.

Our early analysis of this phenomenon indicates a lexical approach is insufficient to fully identify instances of misogynoir online. However, our qualitative examination shows that Racial Gaslighting appears to be a significant problem for the women in our case studies, as well as many of their supporters. Believing Black women in Tech is a theme across all of the case studies. This is connected to Tone Policing in that, if one denies the existence of racial injustice, one can dismiss the anger that arises from it as well. It is also connected to White Centring, in that one can dismiss racial injustice through alternative explanations, the result of which is both misogynistic and racist toward Black women ("white-splaining" racism to those who experience it). While one might observe similar patterns in the way women are treated for discussing sexism, or the ways that Black men may discuss racism, specific stereotypes about Black women create obstacles that neither White women or Black men experience.

The rest of the paper is structured as follows. Section

II describes relevant related work. Section III describes the creation of our lexicon. Section IV describes our analysis of the four high-profile cases of misogynoir in tech. Results of these analysis are presented in Section V. Discussions and conclusions are presented in sections VI and VII respectively. The code, the generated lexicon, the data collected (only tweet IDs following Twitter's publishing guidelines), and the generated annotations is publicly available under https://github.com/kwartengj/Asonam2021.

II. Related Work

Section II-A describes existing literature around misogynoir and provides an analysis of the different categories identified. Section II-B briefly summarises existing work on detecting hateful and abusive speech online, and highlights, how this work contributes to, and advances over, existing efforts.

A. Models of Misogynoir

The basic model of misogynoir is the experience of "gendered racism", but this is difficult to qualify, as it is not simply the sum of its parts. For example, Madden *et al.* [5] conducted a qualitative content analysis of abusive comments received by actress and comedian Leslie Jones, in response to the all-female reboot of the film GhostBusters. The authors identified multiple forms of misogynoir in comments related to her attractiveness to men or perceived "masculine" features, the way her tone and self-boundaries were questioned, and dismissal of the wider context of the abuse she received. This abuse has undertones of both racial and gender stereotypes, but the combined effect is to both dismiss and suppress. Below we describe some of the patterns of misogynoir that have been discussed within the literature and how they recognised in society. Note that these themes can overlap and interact with one another, making a clean distinction between them difficult.

Tokenism: At a general level, tokenism is when an individual is included within an organisation to "represent" a group of people under conditions of continued bias toward that group. A person who is a token may be expected to fulfill colleagues' desires to feel inclusive, or to be all-knowing about issues of diversity and conform (or not) to various stereotypes [6]. This category may be connected with practices such as "diversity branding" in companies, in which the images that are supposed to represent a company's employees or customer base include people of colour, people with disabilities, or other marginalised groups, despite being underrepresented in the company. In technology companies, where women (and particularly Black women) are not highly represented, the danger of Black women being treated like tokens is greater. Thus, we can classify tokenism of Black women in tech as misogynoir. This category is presumed for all of the women in our case studies (see Section IV-A).[3] As tokenism is contextual, and requires background knowledge, it is difficult to identify it online.

[3]https://tinyurl.com/3jf8sf6f, https://tinyurl.com/26p9vspw

White Centring: White Centring is the interpretation of race through white paradigms and interests [7], i.e., when discussions of racism begin to focus on how White people feel being confronted with racism or about racism [8]. Examples include, ignoring other value systems or priorities that are relevant for People of Colour, judging People of Colour against those systems, and making suggestions of how to solve the problem of racism from a White perspective. White Centring is also particularly visible in colourblind or generalised approaches to racial equality, which discount the knowledge of specific groups of people experiencing racism, as well as the features of power and historical circumstances that mediate our interactions [7]. In the field of technology, the pervasive belief is that tech companies are liberal and therefore somehow immune from systemic racism [9]. Coupled with more general experiences of sexism in technology, Black women speaking out about race in tech companies can experience misogynoir as a result of White Centring in a sexist context. All of the women in our case studies reported having experienced sanctions of some sort for speaking about race in their organisations.

Tone Policing: Tone Policing is a mechanism for preserving the status-quo through suppressing expressions of anger in response to injustice [10]. For Black women, Tone Policing is exacerbated by stereotypes of the "angry Black woman" that are ubiquitous in the media and film [5]. One can identify Tone Policing when individuals critique the form and not the content of a serious message about injustice. Calling a person "oversensitive", "hyperbolic", or insinuating this, is Tone Policing. The danger of Tone Policing is that it distracts from the original injustice and creates a secondary problem to "resolve" [11]. As Tone Policing is connected to specific misogynistic and racist stereotypes of Black women, especially in professional contexts, it can be labelled as misogynoir.

Racial Gaslighting: Racial Gaslighting is typically described as using white-centred explanations to undermine evidence of racial inequality specifically, and provide "alternative explanations" for what a Person of Color has experienced as racism. Denying that racism exists, or arguing that Black people "always make it about race", is a form of Racial Gaslighting. It can come in the form of being "unsympathetic to abuse", positioning the recipient of abuse as weak or hyperbolic (connecting with Tone Policing), unable to accept the situation as it is usual or expected in a White interpretation [5]. Because of the additional gendered aspects of women being viewed as emotional or unstable, and Black women as unreasonable or angry, Racial Gaslighting is an even more worrying problem for Black women.

Defensiveness: Defensiveness is a common experience in talking about race and racism with White people [8], [12]. Defensiveness typically appears directly in the form of justification of one's own or another person's behaviours, rejecting any accusations of racism without reflection (potentially a form of White-Centring). As a first response to a racist encounter, justifications indicate a resistance to the narrative that racism is hurtful and common for People of Colour.

Unacknowledged Privilege: Intersectional, Black feminist readings of privilege like Collins [13] and Crenshaw [14] acknowledge a dynamic, interlocking system of oppressions that include aspects of race, gender, class, ability, residential status, religion (or any number of social and demographic features). This allows those who understand this principle to position themselves across many dimensions, and understand their relative advantages and disadvantages. Less sophisticated knowledge around the subject of intersectionality can result in reductive ideas about injustice, in which one's own experience of hardship is given as evidence that privilege does not exist. This appears for many Black women in their interactions with White women around feminism and race [15]. In technology, where White and Black women are struggling for recognition, unacknowledged privilege can make White women poor allies. Unacknowledged privilege is often contained in each of the other forms of misogynoir presented in this section, and is understood as a part of the wider context.

B. Challenges of Detecting Hateful and Abusive Speech

Computational techniques are necessary for both understanding and managing hateful speech online. As a lot of online communication is text based, there is a long history of linguistic computational approaches to analysing online abuse and hateful speech [16].

The content of abuse is however difficult to capture. Specific racial slurs and physical threats are easier to identify with existing techniques because there are clear boundaries around such language (sometimes codified in law). However, most of what people experience on a daily basis is more complex [17], [18]. In addition, online abusers have also adapted, learning to replace racist words with other more benign terms and phrases, to avoid detection [19].

Previous work has tried to capture nuances through delineating certain types of abuse from others using lexicons [20], or providing a set of layered rules for how words interact with each other [21]. Machine Learning techniques, and particularly neural networks, have also developed to automatically identify hate [22]. Although these techniques tend to be more accurate that lexicon-based approaches, they rely on training data, that is often difficult and costly to obtain.

As an initial study into the automated detection of misogynoir, we found a lexical approach to be an appropriate first step, especially given that there is a significant amount of literature that describes the experience and language around misogynoir. To the best of our knowledge, there are currently no existing computational methods and resources that enable the identification of this type of hate automatically.

III. Definitions of Misogynoir

As mentioned previously, one of the key contributions of this work was to create a specific lexicon around misognyoir.

Identifying misogynoir with a lexical approach requires to understand both generalised and context specific terms and expressions. To understand more general experiences, we first experimented with combining terms from existing racism and misogyny detection lexicons from [20]. The combination of these terms did not yield any meaningful results in our dataset so we do not include those results in this paper. Instead, we chose to rely on existing literature to extract the generalised experiences of misogynoir and their linguistic patterns, and combine this with a data-driven set of codes on a subset of our data to understand the more specific experiences of misogynoir. Hybrid approaches of this kind have been shown to improve rigour in exploratory studies [23].

From the scholarly literature, two social science researchers with a background in feminist (and Black feminist studies) were able to identify different forms of misogynoir that are reported by Black women (described in II). Tokenism, White Centring, Tone Policing, Racial Gaslighting, Defensiveness and unacknowledged privilege were prominent themes. Each of those themes were investigated in detail, looking for commonalities in the language used to describe experiences of this form of misogynoir. Several phrases and key sentiments were reported often by Black women, and we included them directly in our lexicon. Then, we conducted an inductive analysis on a sample of 100 tweets about each of our chosen case studies. Our aim was to see what additional specific language is reported in the data as referring to an experience of misogynoir in the context of Black women in tech. We added the terms we extracted from this exercise to the lexicon.

Two annotators who are knowledgeable about the subject worked together to agree on the terms. A future assessment should include individuals who have experienced misogynoir in different contexts. We removed categories where there were not enough terms to describe a very context specific issue, or where too many terms were being used by both those who support and those who do not support the women in our case studies (tokenism and unacknowledged privilege). We arrived at a set of four categories of misogynoir presented below in Table I, and a more general category for messages that are not explicitly one of the other categories. Examples of tweets belonging to each of these categories can be seen in Table II. We created a lexicon for those categories with a total of 254 terms. We also identified three subtypes of supportive messages that users sent in response to the women in our case studies. These were, sharing a personal experience of misogynoir themselves, thanking the woman from the case study for sharing her own experience or generic messages of support (see Table III).

IV. Analysis Approach

We describe in this section the data analysis approach followed. This pipeline is composed by five main phases: (i) selection of case studies, (ii) data collection, (iii) data mapping, i.e., mapping the terms and expressions gathered within the lexicon (see Section III) to the collected data, (iv) data annotation and (v) data filtering. All these difference phases are explained in the subsections below.

A. Case Studies

In this paper, we highlighted four recent cases involving four Black women who were former employees of two large

TABLE I
CATEGORIES OF SIGNIFICANCE INCLUDED IN THE LEXICON

Categories	Definition	Examples	Terms
Tone Policing (TP)	language criticising the form of someone's argument, rather than the content	"not constructive", "complaining about", "whining about"	45
White Centring (WC)	language that seeks to re-contextualise the targets' challenges inside of white culture and values	"why does everything have to be about..., "why didn't she do...",	26
Racial Gaslighting (RG)	language that seeks to downplay or dismiss the role of race in the targets' experience	"reverse racism", "the only race is the human race", "colourblind"	93
Defensiveness (D)	language that talks about calling out bad behaviour as an attack of some sort or an assassination of character	"cancel culture", "block the conversation", "friends who are Black"	39
General (G)	language that more generally refers to racism, sexism or more general support/non-support	"sexism", "Yaaas", "Thank you!"	51

TABLE II
EXAMPLES OF TWEETS FOR EACH CATEGORY

Categories	Example Tweet
Tone Policing (TP)	"I think what you are doing can be called womansplaining your rude and arrogant way of speaking"
White Centring (WC)	"I find it extremely hard to believe Pinterest will send a PI after you. If there are 2 people vying for one promotion, ANY company will 'pit' employees against one other (regardless of their friendship status/ race). Stop blaming your incompetence on race."
Racial Gaslighting (RG)	"From what I can gather, the point is to push the "white people are bad" narrative."
Defensiveness (D)	"So you are saying you've read the email that got her terminated and it was not a firing offense? Or are you just blindly defending another female out of an emotional requirement to defend a perceived social injustice? And you hold a PhD? Fascinating."
General (G)	"wow!"

TABLE III
SUBTYPES OF SUPPORTIVE MESSAGES

Categories	Definition	Examples
Sharing Experiences (E)	users sharing their own experiences of misogynoir as an act of solidarity or allyship	"@company @company .Are some of the most racist companies I worked with. At that time i even had a recruiter say "yeah we know it's a problem but it's a big account for us""
Showing Thanks and Gratitude (T)	users expressing their gratitude toward those sharing their experiences of misogynoir	"Thank you for this", "I'm sorry about this @user and thanks for sharing."
Generic(GR)	More general messages of support	"I am so sorry @user. This is unbelievable. I am speechless."

technology companies, Google and Pinterest. These women told their stories on Twitter about their experiences and why their contracts with these tech companies were terminated.

Dr. Timnit Gebru: Dr. Timnit Gebru was previously a senior researcher in Ethical A.I. at Google. Her work has been instrumental in identifying intersectional challenges in facial recognition [24], gender-based analysis [25] and more recently, on the ethical challenges of using large data models (which ultimately led to her dismissal)[4]. Dr. Timnit Gebru went public with her dismissal on December 3, 2020 via Twitter[5]. In her original messages, she argues that her dismissal follows a pattern of discrimination within Google (and tech more broadly).

April Christina Curley: Since 2014, April Christina Curley has been working as a Diversity Recruiter at Google until she was fired in September 2020. Following stories of Dr. Timnit Gebru's firing triggered her to go public and shared insights about her termination via Twitter on December 21, 2020[6]. April Christina Curley, in her messages, claimed she was faced with active abuse and harassment hence, firing her was a retaliation for her calling Google out.

Ifeoma Ozoma: Ifeoma Ozoma was a second recruit on the Public Policy and Social Impact team at Pinterest for almost two years. Prior to that, she had also served on Google and Facebook's Public Policy unit[7]. At Pinterest, she played an important role in its effort to counter health misinformation on its platform[8]. After seeing Pinterest declare "Solidarity with BLM," Ifeoma Ozoma went public, shared her stories via Twitter on June 15, 2020. Pinterest, she said, had paid little attention to her battle for equal pay and had not addressed an issue in which her personal information was shared with a misogynistic group by a white male colleague[9].

Aerica Shimizu Banks: Aerica Shimizu Banks is an ex-coworker of Ifeoma Ozoma at Pinterest, who was also working in the Public Policy and Social Impact team. She had served on the Patent Policy team at Google before Pinterest[10]. Aerica Shimizu Banks went public about her tenure at Pinterest, and shared her encounter with her former colleague Ifoema Ozoma's Twitter thread. She reported that she was faced with retribution for raising pay disparity and reporting a colleague's derogatory remarks about her ethnicity.[11]

B. Data Collection

We gathered data from the above four identified case studies, from the time they shared their stories until January 2021. This lead to 49,941 tweets being collected. We collected this data using the Twitter API and Snscrape[12], a scraper for

[4]https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

[5]https://twitter.com/timnitGebru/status/1334341991795142667

[6]https://twitter.com/RealAbril/status/1341135819487100928

[7]https://www.linkedin.com/in/ifeomaozoma/

[8]https://techcrunch.com/2019/08/28/pinterest-starts-displaying-information-from-health-organizations-for-searches-related-to-vaccines/

[9]https://twitter.com/IfeomaOzoma/status/1272546213322080258

[10]https://www.aerica.co/home/#aboutme

[11]https://twitter.com/erikashimizu/status/1272547227177713664

[12]https://github.com/JustAnotherArchivist/snscrape

Social Networking Services. Table IV shows a summary of the collected data, including the individuals, the number of tweets gathered for each case study, and the date range for which data was obtained. We collected these tweets using the individuals' Twitter handlers (e.g., @timnitGebru for Dr Timinit Gebru) and their displayed names as keywords. This collection method provided us with a list of tweets for each individual case study which included: (i) replies to the person's posts, (ii) retweets of her posts with additional attached messages by the Twitter users who retweeted them (i.e., quoted retweets), and (iii) original posts by Twitter users naming these individuals. Additionally, to ensure that our dataset captured most of the relevant data surrounding these case studies, we followed their conversations on Twitter. We realised that some Twitter users, rather than quoting the posts authored by our case studies, were posting images of those tweets and commenting on them (e.g, https://twitter.com/jonst0kes/status/1346298617980743680). For completeness we added such posts (100 in total) manually to our dataset.

TABLE IV
NUMBERS OF TWEETS COLLECTED, IDENTIFIED, ANNOTATED AND FILTERED DURING OUR ANALYSIS APPROACH

Case Study	Collected	Mapped	Annotated	Remained	Filtered
A. C. Curley	7634	1656	835	813	22
A. S. Banks	3053	771	625	275	350
I. Ozoma	3360	1170	926	510	416
T. Gebru	35894	5872	987	921	66

C. Data Mapping

Once the data was collected, we matched the collected data against the generated lexicons, to identify the tweets belonging to the above identified categories (see Section III). We first pre-processed the text of each tweet to facilitate the matching by: (i) removing punctuation symbols, (ii) removing numeric symbols, (iii) removing 'RT' and '#' characters, and (iv) transforming the textual content to lowercase. We then proceeded to match the lexicon terms against the data by means of exact term matching, regular expressions, and reducing some of the lexicon terms to their stems to allow for variances (e.g., complaining, complain, etc.) The exact details of the regular expressions and stem variances used to generate this mapping can be found in the provided code.[13] Table IV presents the number of tweets under each category found per case study after conducting this process.

D. Data Annotation

We used our created lexicon to search for the broad themes of misogynoir within our dataset of tweets (White Centring, Defensiveness, etc.). After conducting this mapping process we identified a set of 9,469 Tweets. However, we were not capturing a significant number of misogynistic responses, though we knew from anecdotal statements that each of these women did receive some misogynistic and racist abuse in response to their original statements. To assess the quality of the data mapping process, we conducted a qualitative analysis of the identified tweets by manually annotating a randomly selected subset of approximately 35% of those tweets (a subset of 3,373 tweets), using the codes described in table I. The number of collected, mapped and annotated tweets per use case can be found in Table IV. The annotation process was conducted by two annotators (authors of this paper), with a computed Cohen's Kappa inter-annotator agreement[14] = 0.79 (high agreement). The sample size was determined in advance due to the human cost of manual annotation.

During the annotation process, we discovered that we were capturing many more messages of support than misogynoir, despite our attempts to divide these two groups linguistically in the lexicons. We added another manual coding process to each tweet, to determine it as being in support of the women in our case studies (coded as allyship, "A"), as a potential case of misogynoir (coded as Misogynoir, "M"), or as a potential unclear case (coded as "U").

In the allyship category, we saw 3 prominent types emerging, supporters sharing their own personal experiences of discrimination ("E"), expressions of gratitude and thanks ("T") and more general support for the issue ("GR"). If there was a specific type of misogynoir addressed in the message, we also coded it as White Centring ("WC"), Tone Policing ("TP"), Racial Gaslighting ("RG"), or Defensiveness ("D"). We took these data-driven codes [23] and completed a manual annotation on tweets belonging to the allyship category. We then used the prevalence of these categories to help establish a narrative of how each type of misogynoir or allyship presents itself within the context of our case studies. Results of these analysis are presented in Section V.

E. Data Filtering

During the data annotation process we identified a set of tweets that did not reflect the public's response towards the experiences shared by our case studies. These were tweets, frequently authored by our case studies, responding to comments and retweets, mainly to thank other Twitter users for their support. We decided to remove these tweets from our analysis since our focus is on studying the public response towards self-reported misogynoir. The number of tweets filtered per case study can be seen in Table IV. 2,519 tweets remained after the data filtering. This is the final set of tweets over which results are presented (see Section V).

V. RESULTS

In this section we report the results of both our quantitative and qualitative data analyses together, to be able to contextualise our findings. All the analyses have been conducted over our subset of annotated and filtered data (2,519 tweets).

A. Analysis of the Categories of Misogynoir

Table V shows the number of tweets we identified for each person in our four case studies, by the category of misogynoir

[13]https://github.com/kwartengj/Asonam2021

[14]https://en.wikipedia.org/wiki/Cohen%27s_kappa

that is represented in that tweet. Note that some tweets may belong to more than one category. General messages of support or non-support (G), without much other detail is the largest category. The reasons for this are potentially because of the short format of Twitter posts. The next largest category is Racial Gaslighting (RG). Our qualitative analysis shows that Racial Gaslighting is a significant problem for Black women in tech. This becomes more visible as we look more closely at how the categories play out in the data pertaining to each woman's case in the section below (see Section V-B).

TABLE V
TWEETS PER CATEGORY AND CASE STUDY.

Cat.	A. C. Curley	A. S. Banks	I. Ozoma	T. Gebru
WC	29 (3.3%)	8 (2.6%)	8 (1.4%)	50 (4.9%)
TP	42 (4.7%)	10 (3.2%)	19 (3.3%)	53 (5.2%)
RG	358 (40.3%)	105 (33.9%)	144 (25.3%)	377 (36.9%)
G	428 (48.2%)	181 (58.4%)	387 (68.0%)	450 (44.0%)
D	31 (3.5%)	6 (1.9%)	11 (1.9%)	93 (9.1%)

In table VI, we see the number of tweets that were manually annotated as Allyship, Misogynoir or Unclear. We identified more messages of support (Allyship) in our dataset with our lexicons and qualitative analysis, than messages that could be characterised as misogynoir. It's important to note that this does not mean that misogynoir was not present, since our lexicon may not be capturing all relevant instances.

Black women are reporting direct statements or behaviour they have heard or witnessed, so our lexicon surfaced many examples of supporters talking about misogynoir, rather than misogynoir directed at the women in our case studies as a result of their original statements. In addition, many supporters of the women in our case studies reported instances of abuse, so some examples of misogynoir may have been removed following the platform's regulations. Finally, some reports of misogynoir involved experiences that occurred offline, which are not likely to be captured by this approach, because of the different language used to describe them.

A more fine-grained approach will be necessary to delineate groups of supporters speaking about misogynoir from those engaging in abusive speech that can be characterised as misogynoir (see Discussion). Moreover, the network characteristics of each case study may influence the number of likely supporters or non-supporters in their network. Analysing these aspects is part of our future work.

B. Analysis of Case Studies of Misogynoir

While our methodology may not have captured all examples of misogynoir with our lexical approach, the distributions of the categories still tell a coherent story.

April Christina Curley: We see from VI that April Christina Curley receives more tweets that were annotated as misogynoir than the rest, in particular with Racial Gaslighting and White Centring terms. This may make sense because of her role as a diversity recruiter, which some non-supporters labelled as "reverse racism". She also shared her story after Dr. Timnit Gebru, which some non-supporters viewed as suspect. The expectation that Black women should have capacity to advocate for themselves at all times is another example of misogynoir.

Dr. Timnit Gebru: For Tone Policing, Defensiveness and generic messages, Dr. Timnit Gebru received the most messages that could be considered misogynoir. Dr. Timnit Gebru has 102.4K followers, compared to 23.2K followers for April Christina Curley, 10.6K followers for Ifeoma Ozoma and 4512 followers for Aerica Shimizu Banks. Thus the profile of Dr. Timnit Gebru is more public. It was also the biggest story of the four in the media outside of tech, and Tone Policing played a significant role in discussions online. This was recently renewed as Google fired Margaret Mitchell (who had been sympathetic to Dr. Timnit Gebru) and hired VP Marian Croak (who is a Black woman) to handle their ethics operations. Marian Croak released a message on YouTube about her appointment[15], citing her "diplomatic" approach to fairness in A.I. This was largely read by supporters of Dr. Gebru as a statement Tone Policing the previous ethics team.

Ifeoma Ozoma and Aerica Shimizu Banks: Ifeoma Ozoma and Aerica Shimizu Banks have barely any misogynistic messages in our subset of tweets. This could be because their stories appeared in June of last year and there is more potential that tweets could have been deleted if they broke Twitter's policies. Also, these stories had less press and the individuals' networks are smaller. However, after Pinterest paid a $22.5 million settlement to COO Françoise Brougher, a white woman alleging gender-based discrimination at the company, many individuals showing support for these women are thanking them for their courage in speaking out. This prompted us to look more closely at the category of allyship in the next analysis.

C. Analysis of Allyship and Misogynoir

In looking at expressions of allyship, table VII shows the extent to which those messages showed more general support (GR), whether they thanked the individual for her contribution (T) or whether they shared personal experiences about misogynoir (E). April Christina Curley's data indicated that supporters were sharing experiences across all of the categories of misogynoir, while for the other 3 women, this is more distinct. For Aerica Shimizu Banks, experiences of Defensiveness and Tone Policing appear most prominent. For her colleague Ifeoma Ozoma, White Centring and Defensiveness are the main categories. For Dr.Timnit Gebru, supporters did not appear to be sharing many experiences, but rather validating her own claims with general support. This is most true for Racial Gaslighting. This makes sense as Dr. Timnit Gebru also named her own experience as Racial Gaslighting and many of her supporters are also validating her claim of Racial Gaslighting in their comments.

Below we capture some additional qualitative analysis around the more general experience of misogyny by category.

[15]https://blog.google/technology/ai/marian-croak-responsible-ai/

Tone Policing: Many of the supporters of the women in our case studies said that they felt they could not express anger in appropriate situations. White paradigms of professionalism are evoked to label these women as being overly sensitive, disrespectful or ungrateful (as women employees at high profile tech-companies, with a so-called "great job") to be highlighting racial inequality at their place of work. Allies of the women in our case studies validate their anger, and feel angry too in response.

White Centring: Focusing on White paradigms of acceptable business practice, the types of White Centring that were visible in our data were around what one can expect from an employer (for example, saying that all companies are just after money, it's not that they're racist), who is doing better than who at tech, the potential of researching into the challenges of A.I. disrupting innovation and expecting the women in the case studies to have reported their concerns much earlier. Supporters argue that critiques of when and how Black women speak out ignore the level of risk Black woman in tech assumes, speaking up within a predominantly White, male industry.

Racial Gaslighting: Many supporters of the women in our case studies reported experiencing backlash at their own companies for pointing to injustice (and particularly racial injustice). In tech companies, where the individuals in our case studies were hired for their expertise on diversity (as April Christina Curley, Ifeoma Ozoma and Aerica Shimizu Banks) or ethics (as Dr. Timnit Gebru), the backlash is a direct result of each woman doing the job for which she was hired. In this sense, this category was very much about believing Black women and trusting their experience. It's also about another pattern of silencing Black women, specifically, about systemic racism in systems where Black women are tokenised. This appears to be a significant theme across the case studies, and where a significant portion of harassment was located.

Defensiveness: This category was much harder to spot computationally, because it partially overlaps with nearly every other category.

VI. Discussion

In this paper we have proposed a combination of computational and socio-linguistic methods to semi-automatically analyse the phenomenon of misogynoir online. Our proposed study advances the state of the art by providing novel resources, including a lexicon of terms structured in several categories of significance, a manually annotated dataset of 2,519 tweets, and an in-depth analysis of the public responses towards the self-reporting of four cases of misogynoir. Despite its novelty, it is also important to highlight that this study also presents several limitations.

Our lexical-based approach did not surface as many examples of misogynoir as one would have expected. Experiences of misogynoir, above and beyond the original issue that is spoken about in the case of each woman, continue after each woman goes public with her story. Sometimes, these events happen offline. For example, after Dr. Timnit Gebru was fired, her former employers released a statement[16] in which they presented her termination as her own choice (despite her many public statements to the contrary), which is Racial gaslighting. As discussed previously, the company engaged in a number of public firings and hirings that expressly mention tone in discussing the issue of race, which is Tone Policing. Including these experiences is difficult in a computational analysis. In addition, a handful of very prominent and dedicated harassers have continued to try to contact and threaten the women whose cases we studied, through their personal email, other social media accounts, etc. This has included threats to their person and their safety. Our proposed methodology is currently not able to capture these instances of continued harassment.

There are also multiple comments from low-follower accounts (so-called sock-puppet accounts) that seek to discredit the women in the case studies, or defend the company in question. They sometimes appear to be created for the purposes of harassment and are reported by supporters. While we were analysing our data, we observed a handful of these accounts disappear, presumably spotted by supporters and removed by Twitter for violating site policies. Still, the cohesion of the categories and our ability to find them across the data indicates that we are capturing part of the story.

With an intersectional lens, we know that misogynoir is not the sum of the parts (sexism+racism). Our analysis leaves out other intersectional categories of wealth, ability or sexuality, for example. Making those dimensions more visible would be necessary for a comprehensive approach. In addition, validating models needs to be discussed. This could be done by comparison (e.g. with White women, or Black men), or by triangulation in a larger dataset. These are all avenues for future research.

Being able to automatically identify and analyse misogynoir may help to support the cases of the women we presented here. It is also worth considering how it could harm. There is a chance that trying to automatically detect misogynoir without a very high standard of success could skew the narrative around a very important issue. Misogynoir remains a highly contextual and intimate experience; hence, bringing in direct experience of misogynoir may help us better recognise it.

As a result, big part of our future efforts are directed toward facilitating this process, and generating more fine grained lexicons capturing these experiences. We also aim to explore the use of Machine Learning methods in this context, and their ability to learn from, and identify, this specific type of hate.

It is also important to highlight that, since we evaluated just four use scenarios and on one platform, Twitter, our results cannot be generalised. A higher number of instances of misogynoir may be found on other platforms, and also offline. Exploring how this phenomena manifest in other social networking sites, such as Reddit[17] or Blind[18]; a social network for professionals, is also part of our future work.

[16]https://www.nytimes.com/2020/12/09/technology/timnit-gebru-google-pichai.html

[17]https://www.reddit.com

[18]https://www.teamblind.com

TABLE VI
ALLYSHIP (A), MISOGYNOIR (M) AND UNCLEAR (U) TWEETS PER CATEGORY AND CASE STUDY

	A. C. Curley			**A. S. Banks**			**I. Ozoma**			**T. Gebru**		
Cat.	A	M	U	A	M	U	A	M	U	A	M	U
WC	18 (62%)	4 (14%)	7 (24%)	4 (50%)	0 (0%)	4 (50%)	7 (88%)	0 (0%)	1 (13%)	42 (84%)	4 (8%)	4 (8%)
TP	24 (57%)	4 (10%)	14 (33%)	9 (90%)	0 (0%)	1 (10%)	18 (95%)	0 (0%)	1 (5%)	45 (85%)	5 (9%)	3 (6%)
RG	200 (56%)	48 (13%)	110 (31%)	80 (76%)	1 (1%)	24 (23%)	128 (89%)	4 (3%)	12(8%)	304 (81%)	19 (5%)	54 (14%)
G	383 (89%)	7 (2%)	38 (9%)	157 (87%)	1 (1%)	23 (13%)	357 (92%)	2 (1%)	28 (7%)	385 (86%)	14 (3%)	51 (11%)
D	20 (65%)	3 (10%)	8 (26%)	4 (67%)	0 (0%)	2 (33%)	5 (45%)	1 (9%)	5 (45%)	71 (76%)	14 (15%)	8 (9%)

TABLE VII
EXPERIENCE (E), THANK (T), GENERIC (GR) TWEETS PER CATEGORY AND CASE STUDY

	A. C. Curley			**A. S. Banks**			**I. Ozoma**			**T. Gebru**		
Cat.	E	GR	T	E	GR	T	E	GR	T	E	GR	T
WC	3(17%)	13(72%)	2(11%)	0(0%)	3(75%)	1(25%)	1(14%)	5(71%)	1(14%)	0 (0%)	41(98%)	1(2%)
TP	3(13%)	19(79%)	2(8%)	1(11%)	5(56%)	3(33%)	0(0%)	11(61%)	7(39%)	2(4%)	41(91%)	2(4%)
RG	28(14%)	158(79%)	14(7%)	5(6%)	63(79%)	12(15%)	5(4%)	102(80%)	21(16%)	5(2%)	296(97%)	3(1%)
G	14(4%)	96(25%)	273(71%)	4(3%)	83(53%)	70(45%)	5(1%)	155(43%)	197(55%)	4(1%)	311(81%)	70(18%)
D	2(10%)	13(65%)	5(25%)	1(25%)	3(75%)	0(0%)	1(20%)	4(80%)	0(0%)	1(1%)	70(99%)	0(0%)

Despite the highlighted limitations and open research directions, we believe this work constitutes a first step towards the automatic processing and analysis of online manifestations of misogynoir. While there is ample room for further investigation, we hope that our produced resources (lexicon, annotated dataset) and analyses can incentive the research community to further investigate this phenomenon.

VII. CONCLUSION

In this paper, we analyse the public response in Twitter towards the self-reporting experiences of misogynoir of four Black women in tech. Due to the scarcity of resources to analyse this specific type of hate automatically, we created a novel lexicon capturing different forms of misogynoir using a combined data-driven approach. The different forms of misogynoir (Tone Policing, White Centring, Racial Gaslighting, Defensiveness), and the language around those forms, were identified based on an in-depth review of existing literature. A quantitative and qualitative analysis of the identified case studies have been conducted based on the created lexicon. Our results show how particular categories of misogynoir, like Racial Gaslighting, are more prominent than others, how messages of misogynoir and allyship coexist online, and how the automatically identified online information helps us to capture and analyse public views and responses towards the cases under study.

REFERENCES

1 Bailey, M. and Trudy, "On misogynoir: Citation, erasure, and plagiarism," *Feminist Media Studies*, vol. 18, no. 4, pp. 762–768, 2018.
2 Trudy, "Explanation of misogynoir," *Gradient Lair*, 2014.
3 Epstein, R., Blake, J., and González, T., "Girlhood interrupted: The erasure of black girls' childhood," *Available at SSRN 3000695*, 2017.
4 Tan, Y. C. and Celis, L. E., "Assessing social and intersectional biases in contextualized word representations," in *Advances in Neural Information Processing Systems*, 2019, pp. 13 230–13 241.
5 Madden, S., Janoske, M., Winkler, R. B., and Edgar, A. N., "Mediated misogynoir: Intersecting race and gender in online harassment," in *Mediating misogyny*. Springer, 2018, pp. 71–90.
6 McGee, E. O. and Bentley, L., "The troubled success of black women in stem," *Cognition and Instruction*, vol. 35, no. 4, pp. 265–289, 2017.
7 Mayorga-Gallo, S., "The white-centering logic of diversity ideology," *American Behavioral Scientist*, vol. 63, no. 13, pp. 1789–1809, 2019.
8 Oluo, I., *So you want to talk about race*. Hachette UK, 2019.
9 Noble, S. and Roberts, S., *Technological elites, the meritocracy, and postracial myths in silicon valley*. Duke University Press, 2019.
10 Bailey, A., "On anger, silence, and epistemic injustice," *Royal Institute of Philosophy Supplements*, vol. 84, pp. 93–115, 2018.
11 Nuru, A. K. and Arendt, C. E., "Not so safe a space: Women activists of color's responses to racial microaggressions by white women allies," *Southern Communication Journal*, vol. 84, no. 2, pp. 85–98, 2019.
12 Eddo-Lodge, R., *Why I'm no longer talking to white people about race*. Bloomsbury Publishing, 2020.
13 Collins, P. H., *Intersectionality as critical social theory*. Duke University Press, 2019.
14 Crenshaw, K. W., *On intersectionality: Essential writings*. The New Press, 2017.
15 Bonds, A., "Race and ethnicity ii: White women and the possessive geographies of white supremacy," *Progress in Human Geography*, vol. 44, no. 4, pp. 778–788, 2020.
16 Schmidt, A. and Wiegand, M., "A survey on hate speech detection using natural language processing," in *Proceedings of the fifth international workshop on natural language processing for social media*, 2017, pp. 1–10.
17 Saleem, H. M., Dillon, K. P., Benesch, S., and Ruths, D., "A web of hate: Tackling hateful speech in online social spaces," *arXiv preprint arXiv:1709.10159*, 2017.
18 Gorrell, G., Bakir, M. E., Roberts, I., Greenwood, M. A., and Bontcheva, K., "Which politicians receive abuse? four factors illuminated in the uk general election 2019," *EPJ Data Science*, vol. 9, no. 1, p. 18, 2020.
19 Magu, R., Joshi, K., and Luo, J., "Detecting the hate code on social media," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, 2017.
20 Farrell, T., Fernandez, M., Novotny, J., and Alani, H., "Exploring misogyny across the manosphere in reddit," in *Proceedings of the 10th ACM Conference on Web Science*, 2019, pp. 87–96.
21 Gorrell, G., Bakir, M. E., Greenwood, M. A., Roberts, I., and Bontcheva, K., "Race and religion in online abuse towards uk politicians," *arXiv preprint arXiv:1910.00920*, 2019.
22 Kshirsagar, R., Cukuvac, T., McKeown, K., and McGregor, S., "Predictive embeddings for hate speech detection on twitter," *arXiv preprint arXiv:1809.10644*, 2018.
23 Fereday, J. and Muir-Cochrane, E., "Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development," *International journal of qualitative methods*, vol. 5, no. 1, pp. 80–92, 2006.
24 Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., and Denton, E., "Saving face: Investigating the ethical concerns of facial recognition auditing," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 145–151.
25 Buolamwini, J. and Gebru, T., "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, 2018, pp. 77–91.