Social Network Analysis and Mining (2022) 12:166
https://doi.org/10.1007/s13278-022-00993-7

Misogynoir: challenges in detecting intersectional hate

Joseph Kwarteng[1] · **Serena Coppolino Perfumi**[1] · **Tracie Farrell**[1] · **Aisling Third**[1] · **Miriam Fernandez**[1]

Received: 14 March 2022 / Revised: 23 October 2022 / Accepted: 24 October 2022 / Published online: 9 November 2022

Abstract
"Misogynoir" is a term that refers to the anti-Black forms of misogyny that Black women experience. To explore how current automated hate speech detection approaches perform in detecting this type of hate, we evaluated the performance of two state-of-the-art detection tools, HateSonar and Google's Perspective API, on a balanced dataset of 300 tweets, half of which are examples of misogynoir and half of which are examples of supporting Black women and an imbalanced dataset of 3138 tweets of which 162 tweets are examples of misogynoir and 2976 tweets are examples of allyship tweets. We aim to determine if these tools flag these messages under any of their classifications of hateful speech (e.g. "hate speech", "offensive language", "toxicity" etc.). Close analysis of the classifications and errors shows that current hate speech detection tools are ineffective in detecting misogynoir. They lack sensitivity to context, which is an essential component for misogynoir detection. We found that tweets likely to be classified as hate speech explicitly reference racism or sexism or use profane or aggressive words. Subtle tweets without references to these topics are more challenging to classify. We find that the lack of sensitivity to context may make such tools not only ineffective but potentially harmful to Black women.

Keywords Misogynoir · Hate speech · Social media · Public response · Hate detection · Intersectionality

1 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 (Bailey and Trudy 2018). The term was further developed by Trudy (aka @thetrudz) (Trudy 2014)[1] and the Crunk Feminist Collective[2] to include social or institutional environments (Trudy 2014; Bailey and Trudy 2018). 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 (Epstein et al. 2017). These biases are also visibly encoded in language (Tan and Celis 2019). Understanding misogynoir as a specific type of harm experienced by Black women is important for reshaping industries and fields with low representation.

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 (Madden et al. 2018). This study expands on our previous work (Kwarteng et al. 2021) that analysed the public response in Twitter towards the self-reported experiences of misogynoir of four Black women (case studies) in tech. These Black women were; Dr. Timnit Gebru, April Christiana Curley, Ifeoma Ozoma and Aerica Shimizu Banks. The paper proposed a method to semi-automatically analyse the phenomena of misogynoir online by combining computational and socio-linguistic methodologies. In this

✉ Joseph Kwarteng
joseph.kwarteng@open.ac.uk

Serena Coppolino Perfumi
serena.perfumi@sociology.su.se

Tracie Farrell
tracie.farrell@open.ac.uk

Aisling Third
aisling.third@open.ac.uk

Miriam Fernandez
miriam.fernandez@open.ac.uk

[1] Knowledge Media Institute, The Open University, Walton Hall, Kents Hill, Milton Keynes MK7 6AA, UK

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

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

extended work, we examine and analyse existing methods for automatically detecting hate speech and toxic language and their efficacy in detecting misogynoir. This study aims to: (i) Examine the performance of existing hate speech detection systems in detecting content that can be categorised as misogynoir and (ii) Investigate potential reasons for their performance and opportunities for improvement.

Our contributions can be summarised as follows:

- A newly manually annotated dataset of 2014 Twitter posts combined with our previously annotated dataset from (Kwarteng et al. 2021) of 2519 Twitter posts capturing public responses of misogynoir online (both supportive and non-supportive messages)
- A dataset of 300 Twitter posts multiple-coded as Misogynoir, Allyship, and Tone policing, Racial gaslighting, White centring, Defensiveness and General sampled from the dataset.
- An evaluation of current hate speech detection approaches on our misogynoir dataset.
- An analysis of the challenges and opportunities for understanding misogynoir online.

Our initial examination of this phenomenon reveals that hate speech detection tools are insensitive to detecting instances of misogynoir online. Our qualitative examination shows that the women in our case studies often have their realities of racialised experience questioned (a form of Racial Gaslighting). Believing Black women in Tech is a theme across all of the case studies, in that if one denies the existence of racial injustice, one can dismiss the anger that arises from it as well (a form of Tone Policing). Using alternate explanations, one may dismiss racial injustice, which results in misogyny and racism against Black women ("white-splaining" racism to those who experience it), which is also related to White Centring. 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 performance evaluation of the two state-of-the-art detection methods revealed that HateSonar and Perspective API are ineffective at detecting intersectional hate; misogynoir as they performed poorly. Our qualitative examination of false positives and false negatives revealed that these systems were classifying many instances of tweets containing references to racism, sexism, and profane or aggressive language as hate speech, which makes them more destructive to the Black community and Black women, especially in terms of self-advocacy or the use of African-American English (AAE) which may be inappropriately flagged as racist content. In addition, these systems struggle to identify other subtle types of hate and are insensitive to context, which is a crucial component of misogynoir and intersectional hate.

The rest of the paper is structured as follows. Section 2 describes relevant related work. Section 3 describes the definitions of identified categories and its lexicon. Section 4 describes our analysis approach and how the experiment was conducted. Results of this analysis are presented in Sect. 6. Discussions and conclusions are presented in Sects. 7 and 8 respectively. The code, the newly compiled dataset (only tweet IDs following Twitter's publishing guidelines), and the generated annotations are publicly available under https://github.com/kwartengj/Snam2022.

2 Related work

Section 2.1 describes existing literature around misogynoir and provides an analysis of the different categories identified. Section 2.2 briefly summarises existing work on detecting hateful and abusive speech online and highlights how this work contributes to and advances existing efforts.

2.1 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. 2018) 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 the 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 are recognised in society. Note that these themes can overlap and interact with one another, making a clear distinction between them difficult.

2.1.1 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 fulfil colleagues' desires to feel inclusive or to be all-knowing about issues of diversity and conform (or not) to various stereotypes (McGee and Bentley 2017). 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 general, tokenism is contextual and requires background knowledge, it is difficult to identify it online. 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 the tokenism of Black women in tech as misogynoir. This category is presumed for all of the women in our case studies, so we do not further analyse this category in this article.[3]

2.1.2 White centring

White Centring is the interpretation of race through white paradigms and interests (Mayorga-Gallo 2019), i.e., when discussions of racism begin to focus on how White people feel being confronted with racism or about racism (Oluo 2019). Examples include: ignoring other value systems or priorities that are relevant to 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 (Mayorga-Gallo 2019). In the field of technology, the pervasive belief is that tech companies are liberal and, therefore somehow immune from systemic racism (Noble and Roberts 2019). 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. The combination of Tokenism and White centring is particularly challenging because it places Black women in "other" and "alone" positions. This is why solidarity is important in allyship. Be an ally and show solidarity by actively listening to understand and not responding with scepticism or disbelief when Black women share their stories, but rather by actively advocating for and speaking up for Black women in settings where they are under-represented or unheard. [4]

2.1.3 Tone policing

Tone Policing is a mechanism for preserving the status-quo through suppressing expressions of anger in response to injustice (Bailey 2018). For Black women, Tone Policing is exacerbated by stereotypes of the "angry Black woman" that are ubiquitous in the media and film (Madden et al. 2018). 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" (Nuru and Arendt 2019). As Tone Policing is connected to specific misogynistic and racist stereotypes of Black women, especially in professional contexts, it can be labelled as misogynoir.

2.1.4 Racial gaslighting

Racial Gaslighting is typically described as using white-centred explanations to undermine the 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 (Madden et al. 2018). 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.

2.1.5 Defensiveness

Defensiveness is a common experience in talking about race and racism with White people (Oluo 2019; Eddo-Lodge 2020). 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.

2.1.6 Unacknowledged privilege

Intersectional, Black feminist readings of privilege like (Collins 2019) and (Crenshaw 2017) 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).

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

[4] https://www.forbes.com/sites/hollycorbett/2022/02/22/how-to-be-an-ally-for-black-women-in-the-workplace/?sh=7d49d5fa3123.

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 (Bonds 2020). 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. For this reason, the unacknowledged privilege was not a category of misogynoir that we sought to detect automatically.

2.2 Challenges of detecting hateful and abusive speech

Detecting hate speech is a challenging endeavour as there are several definitional conflicts and variances. According to (MacAvaney et al. 2019), these opposing definitions complicate the assessment of hate speech systems, resulting in datasets derived from disparate sources and capturing disparate information.

Computational techniques are useful 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 (Schmidt and Wiegand 2017).

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 (Saleem et al. 2017; Gorrell et al. 2020). In addition, online abusers have also adapted, learning to replace racist words with other more benign terms and phrases to avoid detection (Magu et al. 2017). Subtle forms of abuse and sarcasm also make the task a challenge. Recent studies (Jurgens et al. 2019; Fortuna and Nunes 2018) that have looked into tackling and proposing subtle hate detection suggest the consideration of making all subtle forms of discrimination, even jokes, as hate speech since they negatively affect some people psychological even though they are considered harmless (Douglass et al. 2016).

Previous work has tried to capture nuances through delineating certain types of abuse from others using lexicons (Farrell et al. 2019), or providing a set of layered rules for how words interact with each other (Gorrell et al. 2019). Machine Learning techniques, and particularly neural networks, have also been developed to automatically identify hate (Kshirsagar et al. 2018). Although these techniques tend to be more accurate than lexicon-based approaches, they rely on training data, which 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 not a significant amount of literature that describes the experience and language around misogynoir (see our earlier work (Kwarteng et al. 2021)). However, this approach did not work as ineffectively as expected since it did not surface as many misogynoir instances as anticipated. 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.

2.3 Intersectional hate detection

A significant body of work that has examined intersectional hate detection has, for the most part, concentrated on addressing intersectional bias in hate speech datasets Maronikolakis et al. (2022); Rankin and Thomas (2020); Kim (2020). (Chandra et al. 2021) combined textual and visual datasets with advanced multimodal deep learning frameworks in order to investigate antisemitism detection. Others focus on individual social identities, specifically either the gender Rodríguez-Sánchez et al. (2020); Park et al. (2018) or the racial Mathew et al. (2021); Sap et al. (2019); Davidson et al. (2019); Waseem (2016) point of view. (Fitzsimons 2022) examined the quantification of intersectional injustice across several demographic groups on Twitter and discovered that the collection of intersectional data is grossly inadequate, and NLP is merely a piece in inherent biases in intersectional hate detections.

In this study, we contribute to previous works in the field by providing a dataset that considers intersectionality within the building process. This dataset is built based on the trigger events of victims who suffer from this hate, i.e. Black women. We then utilised this dataset to evaluate the efficacy of two widely-used detection systems.

3 Definitions of misogynoir terms and expressions

One of the key contributions of our earlier work (Kwarteng et al. 2021) was to create lexicons around misogynoir.

As we have previously noted, this requires both general and context-specific terms and expressions. For generalised content, we relied on existing literature to extract expressions typically related to misogynoir and their linguistic patterns (e.g."whining about race"). To do this, two social science researchers with a background in feminist (and Black feminist) studies mapped terms and phrases to different types

Table 1 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 2 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 3 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."

of misogynoir identified in the literature. Tokenism, White Centring, Tone Policing, Racial Gaslighting, Defensiveness and unacknowledged privilege were prominent themes.

For context specific terms and phrases, we conducted a data-driven, inductive analysis on a subset of 100 tweets about each of our chosen case studies that we categorised as "misogynoir". Terms and phrases here have to do with the specific context of employment at tech companies as a Black woman (e.g. "just do your job" as a response to experiences of racism, or "what does the colour of his skin have to do with it" referring to a specific individual whose behaviour was called out as racist). Hybrid approaches of this kind have been shown to improve rigour in exploratory studies (Fereday and Muir-Cochrane 2006). See our earlier work (Kwarteng et al. 2021) for a more detailed description of the lexicons and how we mapped it to the different types of misogynoir identified.

We arrived at a set of four categories of misogynoir and a more general category for messages that are not explicitly one of the other categories. See Table 1 for the categories, and 2 for examples of tweets belonging to each category. 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 3).

4 Analysis approach

We describe in this section the data analysis approach followed. This pipeline is composed of three main phases: (i) Dataset, (ii) Data annotation and (iii) Hate speech detection tools. All these different phases are explained in the subsections below.

4.1 Dataset

We sampled a total of 2013 tweets from the data gathered in our previous paper (Kwarteng et al. 2021) and removed any duplicated tweets. These tweets had been subjected to the same mapping technique as the reference study and had been labelled by the categories of misogynoir, namely Tone Policing, Racial Gaslighting, Defensiveness, and General (see Sect. 2.1). To evaluate the mapping process's quality, the dataset was then manually annotated using the codes defined in (Kwarteng et al. 2021). Tweets were coded as allyship ("A"), misogynoir ("M"), or an unclear case ("U"). In the case of allyship, tweets were further coded as expressing personal experiences of discrimination "E"), expressions of thankfulness and thanks ("T"), and more generic support for the problem ("GR").

The newly annotated dataset (See Sect. 4.2 for annotation process) was then joined with the analysis dataset of 2519 tweets from the (Kwarteng et al. 2021) article making a total of 4532 tweets for this study. It is worth noting that the 2519 tweets had already been annotated by two annotators (authors of the paper), with a computed Cohen's Kappa inter-annotator agreement[5] value of 0.79 (high agreement) (Kwarteng et al. 2021).

4.2 Data annotation

After extracting the mapped dataset, we conducted an annotation process to label the tweets. We computed inter-annotator agreements by sampling 10% of tweets from the new dataset.

Three annotators participated in the annotation process (authors of this study), which consisted of two stages; first, the annotators individually annotated the dataset as instances of Misogynoir "M", Allyship "A" and Unclear "U" (see Sect. 4.1 for code descriptions). Second, we cross-checked the individual annotations together to ensure a common understanding of the coding principles and consistency of annotation. The objective of debating codes is not to achieve a consensus. It is to identify the points of disagreement and to go deeper into why they exist to offer insights for refining the coding guidelines (Barbour 2001).

Table 4 Numbers of tweets sampled, annotated and filtered during our analysis approach

Labels	Tweets Annotated	Filtered	Remained
Allyship	3862	886	2976
Misogynoir	183	21	162

Note that this data summary does not account for all the potential "U" Unclear tweets

As stated in (Kwarteng et al. 2021), misogynoir is very contextual, and in order to appropriately annotate the dataset, we needed to contextualise and understand the purpose of each tweet. We therefore examined the context of each tweet by using URL links to verify the message and its relations in order to determine its annotation.

Despite this, some tweets still posed significant annotation challenges. For example, as Twitter's policy on offensive and hateful behaviour evolves, tweets and accounts that fall foul of the policy are removed or suspended. These deleted tweets and suspended accounts make it difficult to comprehend the context in which a tweet was authored and even to follow the discussion thread in order to grasp what was said.

Second, annotators found it challenging to label tweets that only contained links or news items relevant to the subject of the case study. For example "Timnit Gebru: Google and big tech are 'institutionally racist' - BBC News https://fook.news/PV5KFj". Upon discussion, we realised that unaltered sharing of news items or URLs without a comment did not clearly distinguish between someone expressing Allyship or Misogynoir, or uninvolved Twitter activity, such as news outlets sharing their own story, or someone sharing a high-profile story to gain impressions. By contrast, if the author of the tweet had added text of their own, this could express a stance. We therefore refined the annotation principles to reflect this.

Additionally, we observed discussions deviating from the case. Thus, users submit derailing tweets underneath threads addressing the narratives of the four case studies. For example, a tweet like "and none of the dinosaurs have the know how to solve the fields deepest and hardest problems western epistemology has a big fat hole in its foundations because of that sexist fascist original bro misogynist aristotle and his brain dead logic". We were not sure whether these were deliberate actions by other users to influence the discourse away from the case studies stories or if they were somehow connected to the discourse.

We computed inter-annotator agreement using Fleiss' kappa[6] from the individual annotation to obtain a kappa value of 0.66 (good agreement). After discussion and

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

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

clarification of the annotation principles, we calculated a new Fleiss' kappa, based on the refined codes, of 0.89. (very good agreement) (Table 4).

For analysis, we removed all the potential cases which were unclear from the data (coded as "U"). The dataset used for analysis only included tweets which were labelled "M" for a potential case of Misogynoir and "A" for a potential case of Allyship (thus tweets that showed support for the women in our case study.)

4.3 Hate speech detection tools

From our previous work (Kwarteng et al. 2021), we learnt that both our case studies' supporters and non-supporters often used the same words and phrases in their tweets. Also, there are currently no computational methods and resources that automatically detect misogynoir. Hence, the need to evaluate the performance of existing hate speech detection tools on a misogynoir dataset to assess its effectiveness in detecting misogynoir as a type of hateful speech. To assess how current hate speech detection tools perform on these lexically comparable classes in detecting misogynoir, we explored two prominent hate speech classifiers: HateSonar and Google's Perspective API.

HateSonar[7] is an open-source automated hate speech detection library for Python based on (Davidson et al. 2017) that classifies text into three categories: (1) Hate speech, (2) Offensive language, and (3) Neither. HateSonar is a Logistic Regression classifier trained on a manually labelled twitter corpus using numerous text features (i.e., TF-IDF of word-grams, sentiment). The classifier is trained on a dataset of 24K tweets that have been labelled by CrowdFlower workers as "Hate Speech", "Offensive Language", or "Neither". Apart from the dataset being extensively utilised as a training dataset in several studies on hate speech detection, including studies by (ElSherief et al. 2018; Davidson et al. 2019; Cao et al. 2020), HateSonar has been used in a number of hate speech detection research studies to evaluate and compare other datasets and classifiers in studies by (Kim et al. 2022; Zannettou et al. 2020).

Google's state-of-the-art hate speech detection tool, dubbed Perspective API,[8] detects potentially harmful textual material, including hate speech. This tool uses machine learning algorithms and a human-curated text corpus to determine each remark's rudeness, contempt, or toxicity. Hundreds of platforms worldwide use Perspective to moderate comments posted by their users—including Reddit, The New York Times, Wall Street Journal, Le Monde, El Pais, Disqus, Coral and OpenWeb.[9] The model was trained using millions of comments from various sources, including online forums like Wikipedia (CC-BY-SA3 licence) and The New York Times. Perspective API has also been used in other studies to review and compare datasets and to identify toxic content (Kumar et al. n.d.; Zannettou et al. 2020; Sap et al. 2019). Perspective's primary attribute is TOXICITY, which scores from 0 to 1, reflecting the expected percentage of annotators who would rate the statement as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion". For instance, if six out of ten raters flagged a remark as toxic, it is labelled with a TOXICITY score of 0.6.

5 Experimental settings

As mentioned in Sect. 4.3, we utilised two existing state-of-the-art hate speech detection tools, namely, HateSonar and Google's Perspective API. We applied these tools to our compiled misogynoir dataset to evaluate their performance in the detection of misogynoir. We evaluated these algorithms over a balanced dataset of 300 randomly sampled tweets which consist of 150 tweets each from the "M" (misogynoir) and "A" (Allyship) labelled tweets. We further conducted a performance evaluation of these algorithms over the entire (imbalanced) dataset. In terms of HateSonar's output labels, we classed "hate_speech" and "offensive_language" labelled tweets with classification confidence ("sonar_confidence") greater than or equal to 0.5 as hateful (potential case of misogynoir in our dataset) and "neither" as not hateful (a potential case of allyship in our dataset). We also applied the same benchmark (greater than or equal to 0.5) for the ("toxicity_score") for the tweets labelled by Google's Perspective API. Note that aside from the classification confidence ("sonar_confidence" and "toxicity_score"), we utilised the standard default parameters of the two hate detection systems [10,11] We generated a classification confusion matrix and a classification assessment report based on the performances of these tools on each dataset.

To assess the performance of the selected hate speech detection tools, we considered the following evaluation metrics: precision, recall, f-measure and accuracy. These measures are computed based on the confusion matrix obtained for each system, which indicates: the number of correctly classified messages as hateful (True positives -TP) or not hateful (True negatives -TN) and the number of incorrectly

[7] https://github.com/Hironsan/HateSonar.

[8] https://www.perspectiveapi.com/.

[9] https://medium.com/jigsaw/10-new-languages-for-perspective-api-8cb0ad599d7c.

[10] https://developers.perspectiveapi.com/s/about-the-api-methods.

[11] https://github.com/Hironsan/HateSonar.

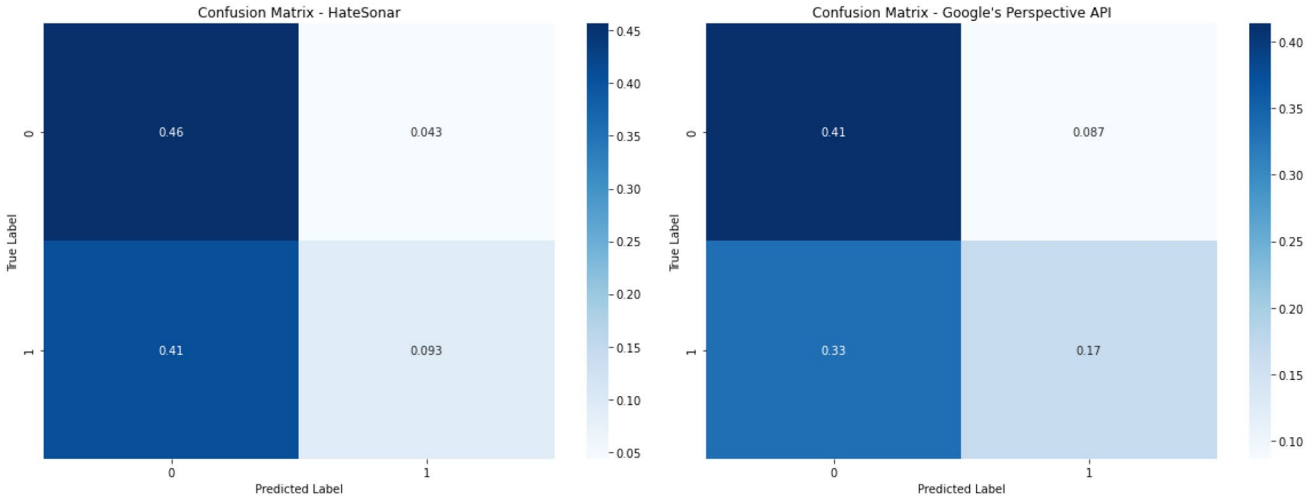

Fig. 1 True versus Predicted Labels on a balanced dataset. (Labels: Misogynoir = 1 and Allyship = 0)

Table 5 Classification report for HateSonar and Google's Perspective API on a balanced dataset

	HateSonar			Google's Perspective API		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Allyship	0.53	0.91	0.67	0.55	0.83	0.66
Misogynoir	0.68	0.19	0.29	0.66	0.33	0.44
Accuracy			0.55			0.58

classified messages, i.e., messages that are classified as hateful when they are not (False positives FP) and messages classified as not hateful when they are actually hateful (False Negatives). See (Seliya et al. 2009) for further details on performance metrics.

6 Results

This section reports the results of our experiments with existing hate speech detection tools and their performance on examples of misogynoir and allyship. As stated in Sect. 5, supporters and non-supporters of our use cases often used the same terms and phrases in their tweets. How do these two types of tools perform on these lexically similar classes? We bring the insights from this study together with our qualitative analyses.

6.1 Balanced dataset

Google's Perspective API outperformed HateSonar on the balanced dataset, with an overall precision of 0.66, recall of 0.33, f1 score of 0.44, and accuracy of 0.58 in identifying misogynoir (see Table 5). According to Fig. 1, 33% of misogynoir tweets are misclassified as not misogynoir, compared to a significantly smaller number of tweets; 17% that are classified as their true label; misogynoir. Nonetheless, approximately 9% of innocuous tweets are incorrectly categorised as misogynoir. As seen in Fig. 1, we can see that both Google's Perspective API and HateSonar find it most challenging to detect misogynoir as only 9.3 and 17% of the 150 tweets labelled misogynoir in the confusion matrix were correctly classified as misogynoir by HateSonar and Perspective API, respectively.

6.2 Imbalanced dataset

Based on Fig. 2, even on an imbalanced dataset, Google's Perspective API outperformed HateSonar with a precision of 0.10, a recall of 0.33 and an f1 score of 0.15 in identifying misogynoir (see Table 6). We can see that the two tools are having difficulty classifying misogynoir, as substantially fewer tweets, 1.7 and 0.96% of the 162 instances of misogynoir were correctly classified by HateSonar and Google's Perspective API, respectively, with the remainder classified incorrectly. While they struggle with misogynoir tweet classification, they appear to perform exceptionally well with non-misogynoir tweet classification, correctly classifying 87 and 80% of the total 2976 instances of non-misogynoir tweets from HateSonar and Google's Perspective API, respectively (see Fig. 2). This is because there are likely more examples of tweets that are not misogynoir

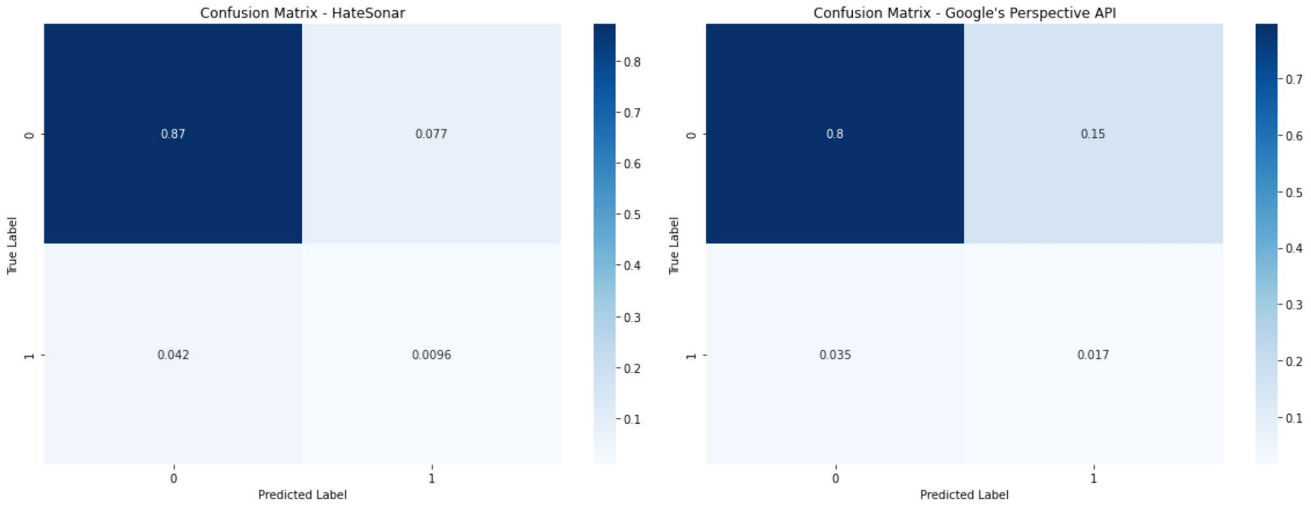

Fig. 2 True versus Predicted Labels on the imbalanced dataset. (Labels: Misogynoir = 1 and Allyship = 0)

Table 6 Classification report for HateSonar and Google's Perspective API on an imbalanced dataset

	HateSonar			Google's Perspective API		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Allyship	0.95	0.92	0.94	0.96	0.84	0.89
Misogynoir	0.11	0.19	0.14	0.10	0.33	0.15

Table 7 HateSonar and Google's Perspective API on misogynoir types

Types	No. of tweets	HateSonar	Google's perspective API
Defensiveness	9	1 (11%)	3 (33%)
General	62	10 (16%)	11 (18%)
Racial gaslighting	53	9 (17%)	27 (57%)
Tone policing	22	7 (32%)	9 (41%)
White centring	16	3 (19%)	3 (19%)

Table 7 shows the correct prediction of HateSonar and Google's Perspective API group by the types of Misogynoir

Table 8 HateSonar and Google's Perspective API on allyship types

Types	No. of tweets	HateSonar	Google's Perspective API
Experience (E)	112	16 (14%)	32 (29%)
Generic (GR)	2003	175 (9%)	366 (18%)
Thanks (T)	861	50 (6%)	82 (10%)

Table 8 shows the correct prediction of HateSonar and Google's Perspective API group by the types of Allyship

in the testing data than tweets that do express misogynoir, which explains why these classifiers perform better on non-misogynoir tweets. See Table 6 for its classification report.

6.3 Analysis of the models of misogynoir and allyship types based on the HateSonar and Perspective API

Table 7 shows the number of misogynoir tweets grouped by the categories of misogynoir and their number of correct classifications made by HateSonar and Google's Perspective API. In general, Racial Gaslighting and Tone Policing are the most significant categories and also with the highest number of tweets accurately classified by the two detection tools. One reason is likely that Racial Gaslighting and Tone Policing do have many lexical clues compared to the other types of misogynoir such as White Centring and Defensiveness which are subtle in nature. Additionally, that category might be overrepresented in both balanced and imbalanced datasets. Racial gaslighting may likely have been the most prevalent type of misogynoir seen by Black women in tech, making it easier to detect. With misogynoir type General, it makes sense that there are an interesting number of tweets correctly classified because its description (see Table 1) indicates that they are languages that include or refer to racism, sexism, or potentially hostile non-support messages. However, subtleties in language continue to pose a barrier to automatic hate speech detection (MacAvaney et al. 2019; Rodríguez-Sánchez et al. 2020) which could be a potential

reason for the poor performance of the two detection tools on misogynoir type Defensiveness and White Centring.

Recent studies (Jurgens et al. 2019; Fortuna and Nunes 2018) that have looked into tackling and proposing subtle hate detection suggest the consideration of making all subtle forms of discrimination, even jokes, as hate speech since they negatively affect some people psychological even though they are considered harmless (Douglass et al. 2016).

In terms of Allyship tweets, Table 8 displays the total number of allyship tweets classified by the categories of allyship and the percentage of tweets classified correctly by HateSonar and Google's Perspective API. As can be seen from the table, these algorithms are not only misclassifying misogynoir tweets but also finding instances of misogynoir inside allyship tweets. These are tweets by authors talking or sharing experiences about misogynoir, which are not direct statements of misogynoir but may include hostile comments. For instance, one tweet said, "i fucking love you you're a genuine girl and now a legend for what you did, fuck these racist ass dumb companies," indicating allyship to one of the women whose experiences of misogynoir in tech companies motivated this study.

6.4 Analysis of misogynoir and allyship based on the HateSonar and Perspective API

To determine why these tweets were misclassified, we now examine the tweets and their anticipated classes in further detail. We observed tweets classified as hate speech by HateSonar, which included occurrences of the term "racist" or "racism" with no clear indication of hate in the sentiment; this was found both in expressions of allyship and misogynoir in the datasets. For example, tweets such as "how is that racist" and "racist detected" are classified as misogynoir with a sonar_confidence of 62 and 71%, respectively. Our research also revealed instances when the term "White" is used in a tweet and is classified as misogynoir. For example, a tweet like; "because you know white women are diversity" is with a sonar_confidence of 62% and "I know. It's so shameful. I won't stop calling out my white people for this shit. It can't get better unless more white people get louder. We have created this mess. It's our responsibility to clean it up, even though we cause irreparable damage, still we must try" is with a sonar_confidence of 69%. These tools may be flagging anything that has a racial marker as hateful. In a Black feminist interpretation of racism, power is an essential feature in determining what is ultimately racism. Therefore, general approaches which view all racial markers as hateful will flag Black women's sense-making activities around White allyship as hateful speech.

In the case of Google's Perspective API, the tweets likely to be labelled as misogynoir are those that include swear or curse words or other profane language. For instance, "wtf an accent is not a disability, and in any case, it's illegal for them to ask you about disabilities also fuck them for insulting our home town", and "stop with the angry black woman bullshit" are scored 89 and 95% toxic respectively, and are in turn classed as misogynoir. We argue strongly that classifying strongly-worded statements that call out racism as toxic is problematic, which is computational tone-policing.

While HateSonar and Perspective API are effective at recognising tweets containing anti-black racism, hostile, sexist and swear slurs, which may constitute misogynoir, it is less effective at detecting nuanced types of misogynoir and hate speech in general, as observed by (Davidson et al. 2017; Nobata et al. 2016). For example, "you got fired get over it" is misogynoir in the sense that it contains elements of Racial Gaslighting for dismissing a Black woman's experience of racism, as well as White-centrism for deciding how someone (a Black woman) should deal with an experience of racism. These tweets are misclassified as not misogynoir, possibly because they contain no racist, sexist, or profane terms or make references to these topics.

There is a strong implication here for the Black community and women in general that they will be labelled as hateful for speaking out against racism, and sexism or making references to experiences of misogynoir that contain these hateful slurs by these detection systems. This finding is consistent with earlier research demonstrating that tweets in the African American English (AAE) dialect are up to two times more likely than other tweets to be labelled as hateful/offensive/toxic (Sap et al. 2019). We believe that this is a mere reflection of what occurs in society. Again, computational tone policing, as mentioned earlier.

6.5 Analysis of Google's Perspective API attributes

Our previous results showed that Google's Perspective API outperformed HateSonar in detecting misogynoir as hate speech. Given this, we conducted a more in-depth experiment to investigate the other attributes of the Perspective API. The Perspective API supports six production attributes namely; Toxicity, Severe_Toxicity, Identity Attack, Insult, Profanity and Threat .[12] These attributes have been tested across several domains and trained using a large volume of human-annotated text. This section summarises our results, assessing five of these attributes and their effectiveness in recognising misogynoir as hate speech. Two tweets from the dataset were excluded due to containing languages not being supported by some of the Perspective API attributes.

[12] https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages.

Table 9 Classification report for the attributes of Google's Perspective API on an balanced dataset

Attributes	Precision	Recall	F1-Score	Accuracy
Severe toxicity	0.67	0.12	0.20	0.53
Identity attack	0.66	0.31	0.42	0.57
Insult	0.70	0.29	0.41	0.58
Profanity	0.62	0.10	0.17	0.52
Threat	0.74	0.11	0.20	0.54

Table 10 Classification report for the attributes of Google's Perspective API on an imbalanced dataset

Attributes	Precision	Recall	F1-Score	Accuracy
Severe toxicity	0.11	0.12	0.11	0.90
Identity attack	0.09	0.32	0.14	0.79
Insult	0.11	0.29	0.16	0.84
Profanity	0.08	0.10	0.09	0.89
Threat	0.01	0.13	0.11	0.90

We produced a confusion matrix of the classification results based on the attributes on our balanced (see Fig. 3) and imbalanced (see Fig. 4) datasets.

Across the balanced and imbalanced dataset, attributes such as IDENTITY ATTACK and INSULT outperformed the other Perspective API attributes in classifying misogynoir messages as hate; see Tables 9 and 10 for their classification report. We discovered that the tweets correctly identified by IDENTITY ATTACK have a combination of the word "you", "people" and a racial identifier such as "white" or "black" in the tweets. We also noticed a mix of phrases such as "toxic", racial markers such as "white" or "black", and conversations referencing the term "racist" in the correctly identified INSULT tweets.

7 Discussion

In this paper, we built on our previous paper (Kwarteng et al. 2021) that analysed the public response on Twitter towards the self-reported experiences of misogynoir of four Black women in tech. That study proposed a combination of computational and socio-linguistic methods to analyse the phenomenon of misogynoir online semi-automatically. We extended this work by examining existing approaches for detecting hate speech automatically and assessing their effectiveness in detecting misogynoir. On our dataset of 3,138 tweets labelled misogynoir and allyship, we proposed a study to evaluate the performance of two popular detection systems, HateSonar and Google's Perspective API.

Our experiment revealed that existing hate speech detection tools are ineffective at detecting this type of hate, misogynoir. They are not sensitive enough to contexts, such as; the individuals involved, their particular circumstances, URL links or images associated with the text, or an article being commented on and the broader discourse around the issue, which can result in such tools identifying sense-making activities around allyship or experiences of racism as harmful or hateful speech. We argue that this is a form of computational White-Centring and Racial Gaslighting (See Sect. 2.1). In our use cases, the additional context of tokenism in tech is not able to be taken into consideration in automated techniques. In addition, they potentially rely too heavily on explicit language to determine harm, which exacerbates the above and can amount, in the worst cases, to computational tone policing.

We observed that these detection tools were picking up tweets making references to racism and sexism and including swearing or profane terms. This finding is consistent with past research that racism is a more pervasive form of hate speech (Silva et al. 2016). This may explain why these algorithms are identifying some instances of such tweets, as there is a clear boundary surrounding such language. We saw forms of misogynoir such as Racial Gaslighting and Tone Policing occurring in tweets detected to be making references to racism, sexism and swear words since they include more lexical cues than the other forms.

As mentioned in Sect. 6.3, subtle forms of misogynoir like White Centring and Defensiveness are challenging for these current approaches to identify. Subtle hate is still a challenge to detect (Rodríguez-Sánchez et al. 2020), and these classifiers are not context-aware, which might be a possible cause. For example, the dataset used to train HateSonar was built using a Hatebase[13] hate speech vocabulary, including commonly used terms and phrases on the internet, which is likely to create non-representative training data with other nuances of hate uncounted for. Additionally, most training datasets for research on hate speech detection depend heavily on crowd-sourced raters, who may lack knowledge of misogynoir. As discussed in (Kwarteng et al. 2021), context is essential to misogynoir identification; hence, understanding the context and experiences of misogynoir can assist in its detection.

Our research has some limitations. First, the classifiers' labels do not match the labels in the dataset. For instance, HateSonar classifies tweets into hate speech, offensive language and neither. To use HateSonar, we treated both hate speech and offensive language as potential cases of misogynoir, which might not be an accurate representation of what constitutes misogynoir.

[13] https://hatebase.org/.

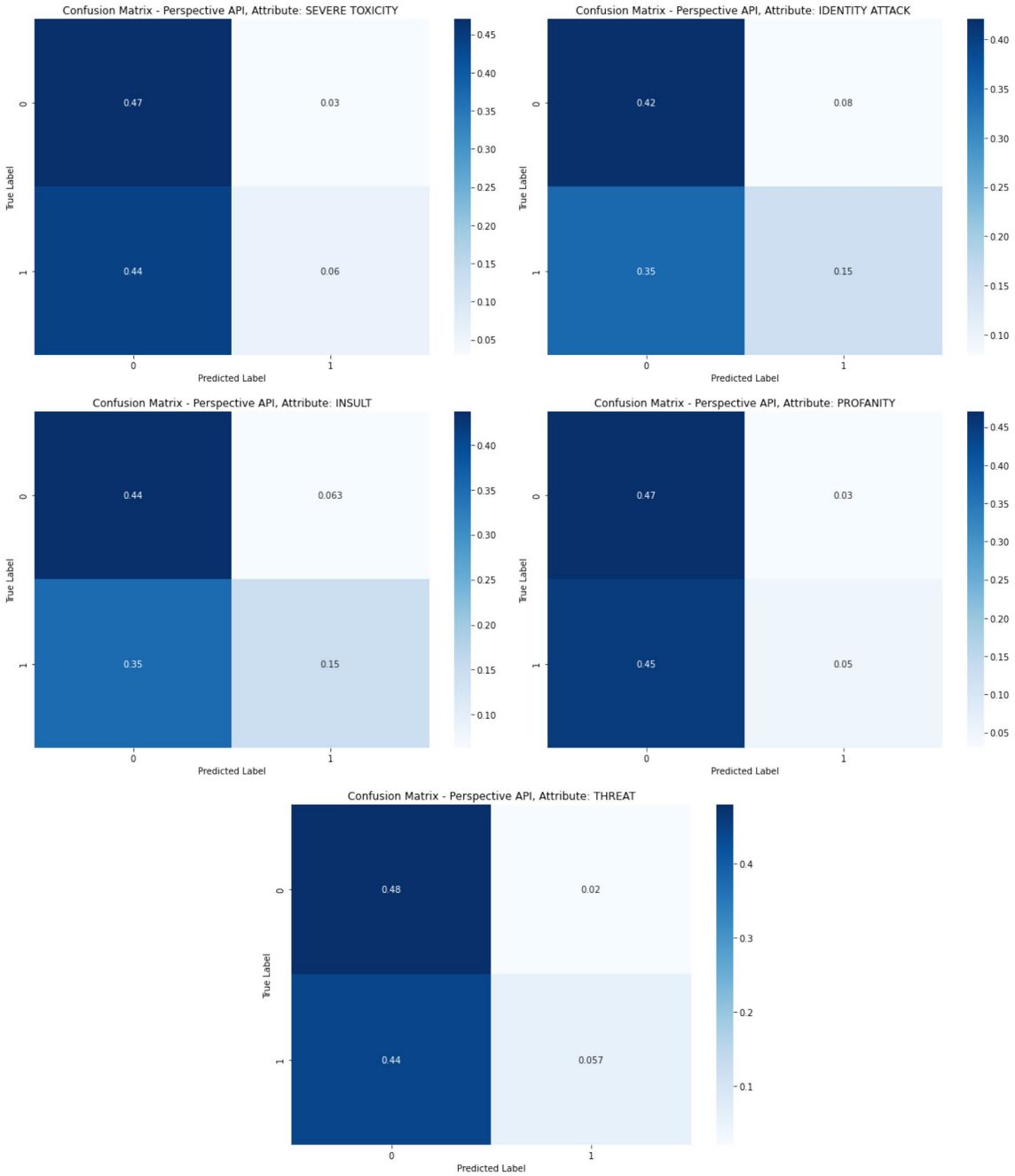

Fig. 3 True versus Predicted Labels of Google's Perspective API attributes on the balanced dataset. (Labels: Misogynoir = 1 and Allyship = 0)

Bear in mind that detecting hate speech is an open research subject, and no classifier can identify all types and forms of hate speech to the best of our knowledge. We also plan to experiment with a combination of methods to arrive at a more robust approach—using computational and qualitative methods to explore diversity in dataset curation, how the involvement of the target of this hate might influence the annotation processes, and how to make these systems context-aware.

As a future scope, further work is needed to understand and detect the intersection of two or more social identities to ensure the social equality and non-discriminatory nature of these existing hate speech detection systems. Future approaches will focus on automated detection that will have to be context-aware, sensitive to issues of power and privilege and reduce the harmful impact of false classification on Black women—addressing bias and under-representation of diversity or targets of hateful content in the training data

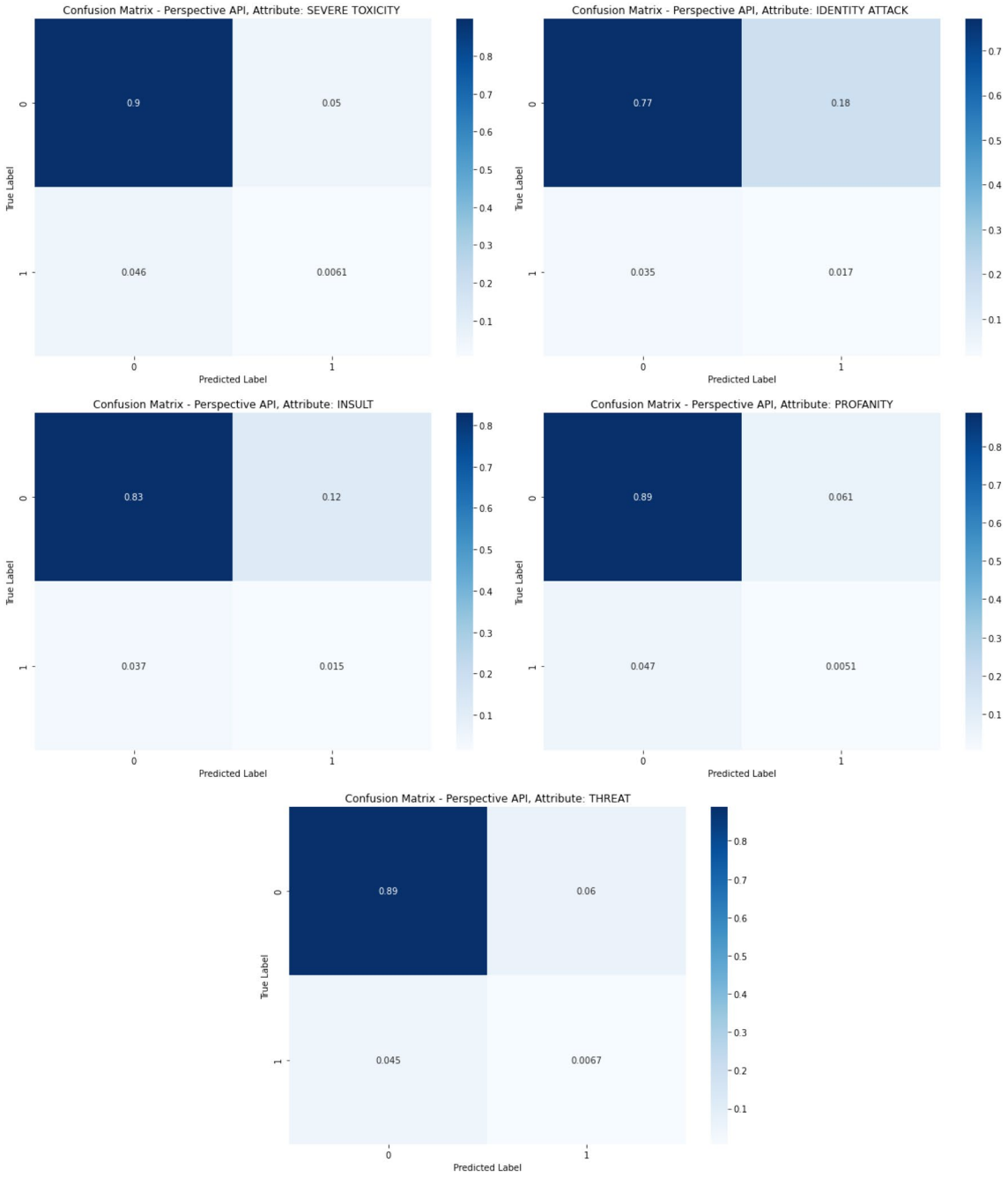

Fig. 4 True versus Predicted Labels of Google's Perspective API attributes on the imbalanced dataset. (Labels: Misogynoir = 1 and Allyship = 0)

and annotation processes. Some of the datasets accessible or used to train detection algorithms, for instance (Davidson et al. 2017; Waseem and Hovy 2016; Gomez et al. 2019), are sampled using an ad hoc collection of phrases or crowd-sourced dictionaries of hateful expressions .[14] This makes them more likely to provide an unrepresentative sample or training data that may not adequately reflect minority communities.

[14] https://hatebase.org/.

8 Conclusion

In this paper, we generated a new dataset that can be used for future research on misogynoir detection on social media. The dataset for this study consists of 162 misogynoir and 2976 allyship tweets carefully labelled and agreed upon by annotators. We evaluated the performance of two existing state-of-the-art hate speech detection systems HateSonar and Google's Perspective API, on our misogynoir dataset in order to determine their effectiveness in classifying misogynoir as hateful speech.

In our performance evaluation of the two state-of-the-art detection systems, we observed that they were ineffective at detecting misogynoir. They performed poorly at detecting many instances of misogynoir as toxic or hateful. Despite their inability to detect nuanced kinds of hate, the Perspective API performed better than HateSonar, which could be due to the high volume of data from different platforms that Perspective API was trained on and its data gathering process as to the 24K data gathered using a set of ad hoc hate speech terms by HateSonar (See Sect. 4.3). Our qualitative analysis of the false positives and false negatives of the predictions done by HateSonar and Perspective API indicates that, in cases where they detect misogynoir correctly, they identify tweets that make explicit references to racism or sexism or use profane or aggressive words. This means that Black women talking about racism online, particularly when they are doing so in a forceful way, will also likely be classified as engaging in hateful speech. This can have a chilling effect on Black women's self-advocacy. It also amounts to computational tone-policing, which mirrors experiences of misogynoir throughout society.

This study demonstrates that further effort is required to enhance all-purpose hate speech detection algorithms in order to address more nuanced and subtle kinds of hatred, such as intersectional hate.

Author Contributions All authors contributed to the study's conception and design. Material preparation, data collection and analysis were performed by Joseph Kwarteng, Data Annotation by Serena Coppolino Perfumi, Tracie Farrell, Aisling Third and Miriam Fernandez. The first draft of the manuscript was written by Joseph Kwarteng, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data Availability The datasets generated during and/or analysed during the current study are available in the GitHub repository, https://github.com/kwartengj/Snam2022.

Declarations

Conflict of interest The authors declare that no commercial or financial ties existed that might be interpreted as a possible conflict of interest while conducting the study.

References

Bailey A (2018) On anger, silence, and epistemic injustice. Royal Instit Philos Suppl 84:93–115

Bailey M, Trudy R (2018) On misogynoir: citation, erasure, and plagiarism. Fem Med Stud 18(4):762–768

Barbour RS (2001) Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? BMJ 322(7294):1115–1117

Bonds A (2020) Race and ethnicity ii: white women and the possessive geographies of white supremacy. Prog Hum Geogr 44(4):778–788

Cao R, Lee RK-W, Hoang T-A (2020) Deephate: Hate speech detection via multi-faceted text representations. In: 12th ACM conference on web science pp. 11–20

Chandra M, Pailla D, Bhatia H, Sanchawala A, Gupta M, Shrivastava M, Kumaraguru P (2021) "Subverting the Jewtocracy": online antisemitism detection using multimodal deep learning. pp. 148–157. Association for Computing Machinery (ACM). Retrieved from https://arxiv-org.libezproxy.open.ac.uk/abs/2104.05947v3https://doi.org/10.1145/3447535.3462502

Collins PH (2019) Intersectionality as critical social theory. Duke University Press, USA

Crenshaw KW (2017) On intersectionality: essential writings. The New Press

Davidson T, Bhattacharya D, Weber I (2019) Racial bias in hate speech and abusive language detection datasets. arXiv preprint arXiv:1905.12516

Davidson T, Warmsley D, Macy M, Weber I (2017) Automated hate speech detection and the problem of offensive language. 11(1)

Douglass S, Mirpuri S, English D, Yip T (2016) They were just making jokes: ethnic/racial teasing and discrimination among adolescents. Cult Divers Ethnic Minority Psychol 22(1):69

Eddo-Lodge R (2020) Why i'm no longer talking to white people about race. Bloomsbury Publishing, UK

ElSherief M, Kulkarni V, Nguyen D, Wang WY, Belding E (2018) Hate lingo: A target-based linguistic analysis of hate speech in social media. In: Proceedings of the international AAAI conference on web and social media Vol. 12

Epstein R, Blake J, González T (2017) Girlhood interrupted: The erasure of black girls' childhood. Available at SSRN 3000695

Farrell T, Fernandez M, Novotny J, Alani H (2019) Exploring misogyny across the manosphere in reddit. In: Proceedings of the 10th ACM conference on web science pp. 87–96

Fereday J, Muir-Cochrane E (2006) Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. Int J Qual Methods 5(1):80–92

Fitzsimons A (2022) Intersectional identities and machine learning: illuminating language biases in twitter algorithms. Hicss pp. 1–10)

Fortuna P, Nunes S (2018) A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR) 51(4):1–30

Gomez R, Gibert J, Gómez L, Karatzas D (2019) Exploring hate speech detection in multimodal publications. CoRR, abs/1910.03814 . Retrieved from http://arxiv.org/abs/1910.03814arXiv:1910.03814

Gorrell G, Bakir ME, Greenwood MA, Roberts I, Bontcheva K (2019) Race and religion in online abuse towards UK politicians. arXiv preprint arXiv:1910.00920

Gorrell G, Bakir ME, Roberts I, Greenwood MA, Bontcheva K (2020) Which politicians receive abuse? four factors illuminated in the UK general election 2019. EPJ Data Sci 9(1):18

Jurgens D, Chandrasekharan E, Hemphill L (2019) A just and comprehensive strategy for using NLP to address online abuse. arXiv preprint arXiv:1906.01738

Kim J, Wohn DY, Cha M (2022) Understanding and identifying the use of emotes in toxic chat on twitch. Online Social Netw Media 27:100180

Kim JY, Ortiz C, Nam S, Santiago S, Datta V (2020) Intersectional bias in hate speech and abusive language datasets. Retrieved from https://arxiv-org.libezproxy.open.ac.uk/abs/2005.05921v3arXiv:2005.05921, https://doi.org/10.48550/arxiv.2005.05921

Kshirsagar R, Cukuvac T, McKeown K, McGregor S (2018) Predictive embeddings for hate speech detection on twitter. arXiv preprint arXiv:1809.10644

Kumar D, Mason J, Bailey M, Gage P, Consolvo KS, Bursztein E, Thomas K (n.d.) This paper is included in the Proceedings of the 17th symposium on usable privacy and security. Designing toxic content classification for a diversity of perspectives designing toxic content classification for a diversity of perspectives. Retrieved from https://data.esrg.stanford.edu/study/toxicity-perspectives

Kwarteng J, Perfumi SC, Farrell T, Fernandez M (2021) Misogynoir: public online response towards self-reported misogynoir. In: Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining pp. 228–235

MacAvaney S, Yao H-R, Yang E, Russell K, Goharian N, Frieder O (2019) Hate speech detection: challenges and solutions. PLoS ONE 14(8):e0221152

Madden S, Janoske M, Winkler RB, Edgar AN (2018) Mediated misogynoir: intersecting race and gender in online harassment. Mediating misogyny, Springer, pp. 71–90

Magu R, Joshi K, Luo J (2017) Detecting the hate code on social media. In: Proceedings of the international AAAI conference on web and social media Vol. 11

Maronikolakis A, Baader P, Schütze H (2022) Analyzing hate speech data along racial, gender and intersectional axes. pp. 1– 7. Association for Computational Linguistics (ACL). Retrieved from https://arxiv-org.libezproxy.open.ac.uk/abs/2205.06621v2, https://doi.org/10.18653/v1/2022.gebnlp-1.1

Mathew B, Saha P, Yimam SM, Biemann C, Goyal P, Mukherjee A (2021) HateXplain: A benchmark dataset for explainable hate speech detection. In: 35th AAAI conference on artificial intelligence, AAAI 2021 Vol. 17A, pp. 14867–14875. Retrieved from https://github.com/punyajoy/HateXplain

Mayorga-Gallo S (2019) The white-centering logic of diversity ideology. Am Behav Sci 63(13):1789–1809

McGee EO, Bentley L (2017) The troubled success of black women in stem. Cogn Instr 35(4):265–289

Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y (2016) Abusive language detection in online user content. In: Proceedings of the 25th international conference on world wide web pp. 145–153

Noble S, Roberts S (2019) Technological elites, the meritocracy, and postracial myths in silicon valley. Duke University Press, UK

Nuru AK, Arendt CE (2019) Not so safe a space: women activists of color's responses to racial microaggressions by white women allies. South Commun J 84(2):85–98

Oluo I (2019) So you want to talk about race. Hachette, UK

Park JH, Shin J, Fung P (2018) Reducing gender bias in abusive language detection. pp. 2799–2804. Retrieved from arXiv:abs/1808.07231

Rankin YA, Thomas JO (2020) The intersectional experiences of black women in computing. In: Proceedings of the 51st ACM technical symposium on computer science education pp. 199–205

Rodríguez-Sánchez F, Carrillo-de Albornoz J, Plaza L (2020) Automatic classification of sexism in social networks: an empirical study on twitter data. IEEE Access 8:219563–219576

Saleem HM, Dillon KP, Benesch S, Ruths D (2017) A web of hate: Tackling hateful speech in online social spaces. arXiv preprint arXiv:1709.10159

Sap M, Card D, Gabriel S, Choi Y, Smith NA (2019) The risk of racial bias in hate speech detection. In: Proceedings of the 57th annual meeting of the association for computational linguistics pp. 1668–1678

Schmidt A, Wiegand M (2017) A survey on hate speech detection using natural language processing. In: Proceedings of the 5th international workshop on natural language processing for social media pp. 1–10

Seliya N, Khoshgoftaar TM, Van Hulse J (2009) A study on the relationships of classifier performance metrics. In: 2009 21st IEEE international conference on tools with artificial intelligence pp. 59–66

Silva L, Mondal M, Correa D, Benevenuto F, Weber I (2016) Analyzing the targets of hate in online social media. In: 10th international AAAI conference on web and social media

Tan YC, Celis LE (2019) Assessing social and intersectional biases in contextualized word representations. Advances in neural information processing systems pp. 13230–13241

Trudy (2014) Explanation of misogynoir. Gradient Lair

Waseem Z (2016) Are You a Racist or Am I Seeing Things? annotator influence on hate speech detection on twitter. In: NLP + CSS 2016 - emnlp 2016 workshop on natural language processing and computational social science, proceedings of the workshop pp. 138–142. Retrieved from www.spacy.io, https://doi.org/10.18653/v1/w16-5618

Waseem Z, Hovy D (2016) Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL student research workshop pp. 88–93. Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/N16-2013, https://doi.org/10.18653/v1/N16-2013

Zannettou S, ElSherief M, Belding E, Nilizadeh S, Stringhini G (2020) Measuring and characterizing hate speech on news websites. In: 12th ACM conference on web science pp. 125–134