Linguistically Differentiating Acts and Recalls of Racial Microaggressions on Social Media

Uma S. Gunturi

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Masters of Science

in

Computer Science and Applications

Eugenia H. Rho, Chair

Edward Fox

Ismini Lourentzou

May 3, 2023

Blacksburg, Virginia

Keywords: Human Centered Computing, Natural Language Processing, Race and Ethnicity

Linguistically Differentiating Acts and Recalls of Racial Microaggressions on Social Media

Uma S. Gunturi

(ABSTRACT)

Experiences of interpersonal racism persist as a prevalent reality for BIPOC (Black, Indigenous, People of Color) in the United States. One form of racism that often goes unnoticed is racial microaggressions. These are subtle acts of racism that leave victims questioning the intent of the aggressor. The line of offense is often unclear, as these acts are disguised through humor or seemingly harmless intentions. In this study, we analyze the language used in online racial microaggressions (“Acts”) and compare it to personal narratives recounting experiences of such aggressions (“Recalls”) by Black social media users. We curated a corpus of acts and recalls from social media discussions on platforms like Reddit and Tumblr. Additionally, we collaborated with Black participants in a workshop to hand-annotate and verify the corpus. Using natural language processing techniques and qualitative analysis, we examine the language underlying acts and recalls of racial microaggressions. Our goal is to understand the lexical patterns that differentiate the two in the context of racism in the U.S. Our findings indicate that neural language models can accurately classify acts and recalls, revealing contextual words that associate Blacks with objects that perpetuate negative stereotypes. We also observe overlapping linguistic signatures between acts and recalls, serving different purposes, which have implications for current challenges in social media content moderation systems.

Linguistically Differentiating Acts and Recalls of Racial Microaggressions on Social Media

Uma S. Gunturi

(GENERAL AUDIENCE ABSTRACT)

Racial Microaggressions are expressions of human biases that are subtly disguised. The differences in language and themes used in instances of Racial Microaggressions ("Acts") and the discussions addressing them ("Recalls") on online communities have made it difficult for researchers to automatically quantify and extract these differences. In this study, we introduce a tool that can effectively distinguish acts and recalls of microaggressions. We utilize Natural Language Processing techniques to classify and identify key distinctions in language usage and themes. Additionally, we employ qualitative methods and engage in workshop discussions with Black participants to interpret the classification results. Our findings reveal common linguistic patterns between acts and recalls that serve opposing purposes. Acts tend to stereotype and degrade Black people, while recalls seek to portray their discomfort and seek validation for their experiences. These findings highlight why recalls are often considered toxic in online communities. This also represents an initial step towards creating a socio-technical system that safeguards the experiences of racial minority groups.

To my parents, Ravi and Vasanta, and sister Tara, for all your love and support.

To my adviser, Eugenia, for your unrelenting guidance and confidence

In loving memory of Raghu Babayya.

Acknowledgments

There are many people to whom I am grateful and without whom the thesis would have been almost impossible to write (much less finish)[1]. Firstly, I am immensely grateful to my advisor, Dr. Eugenia Rho. There is truly no way for me to express my gratitude for her in a few words. Nonetheless, I must thank her for putting up with my constant flow of ideas and for having the patience to allow me to learn from my own errant ideas and mistakes. She occasionally (quite kindly) reminded me not to be the hammer desperately searching for a nail when motivating critical research problems. One can say, she has set an unimaginably high standard for my future advisors and myself to match. I thank Dr. Edward Fox and Dr. Ismini Lourentzou for readily agreeing to serve as members on the committee, for diligently reading drafts of this thesis, and for providing very helpful feedback after my defense. This thesis would never have materialized without my SAIL Research group friends and lab mates, Xiaohan Ding, Anisha Kumar, Rohan Leekha, Kaike Ping and Buse Carik. It was an honor to be a part of such an electric and bright group, in which everyone supported each other and helped me grow in many ways over the last two years. Special thanks to Xiaohan Ding and Anisha Kumar for their integral contributions to this work. I am grateful to my amazing

[1]The structure and wording of this acknowledgements section draw inspiration from the dissertations of numerous PhD students, research scientists, and professors whose works have served as a source of inspiration throughout my academic journey. I am grateful for their contributions, and while I have employed similar words in expressing my gratitude, the sentiments expressed here are sincere and specific to my own experiences.

mentors from my summer internship at IBM Research- Alexander Gray, Achille Fokoue and Tengfei Ma. I'd also like to thank fellow interns and friends, Muneeza Azmat and Vishal Pallagani, who made my time there memorable. During my time as an undergrad intern at the IIIT Delhi MIDAS group, I learned a great deal from my talented mentors Debanjan Mahta and Rajiv Ratn Shah, as well as my collaborators Laiba Mehnaz and Isabelle Lee. I am highly grateful to them for being largely responsible for piquing my interest in the field of code-switching and computational social science. I am also grateful for the many friends who knowingly or unknowingly aided me in making critical decisions over the years. Devyanshu Shukla nurtured my passion for machine learning very early on. He might not know it, but without him and the environment he fostered during my undergraduate career, it is hard to imagine being where I am now. I would like to thank Anisha Kumar for wearing multiple hats throughout this journey. I will immensely miss the fun drives, dinners, brainstorming sessions, brunches, and most importantly deep conversations with her. Finally, I am grateful for my supportive friends- Hemanth Krishna Gudavalli, Shubhi Gupta, Harshit Dokala and Kevin Gautham Radja. On a side note, thanks to my piano teacher Galla Matekova for letting me find my roots and keeping me grounded. I am grateful to all of the students of CS 2114 who I had the honor of teaching since Fall 2021 for helping me grow as a teacher. Special thanks to Derek Haaq and all the co-TAs for allowing me to practice how to make even the most complex concepts digestible to a beginner. I acknowledge the support from Virginia Tech and am honored to receive the Paul E. Torgerson Research Excellence award recognizing this work. A special thank you to my partner Sanath for supporting me in and out of work and adding meaning to my life outside of work. I would also like to thank my sister, Tara, for being my hero in every sense of word and for being my first mentor in every aspect of my thesis journey. The final thank you must go to my parents, Ravi Srinivas Gunturi and Vasanta Sai Bhamidipati, for their patience in allowing me to explore what made me happy and their enduring encouragement in allowing me to forge my own path. I

am extremely fortunate that they made it easy for me to dream freely and to pursue those dreams.

Contents

List of Tables

Chapter 1

Introduction

Experiences of interpersonal racism - whether implicit or explicit - are still a regular part of life for most Black individuals living in the United States. While overt forms of racism against Blacks may have subsided compared to the decades past [114], race scholars argue that racism in modern society has not gone away; rather, it has morphed into more implicit and covert forms of expressions and subconscious acts that manifest in everyday life [127]. This *modern racism*, which is also referred to as symbolic racism or racial microaggressions, is often "highly disguised, invisible, and takes on subtle forms that lie outside the level of conscious awareness" [127, 144]. In this work, we examine the linguistic signature of online racial microaggressions and how it differs from that of personal narratives recalling experiences of such aggressions shared by Black social media users.

Racial microaggressions are subtle acts of racism that often leave the victims questioning the intent of the aggressor, as the line of offense is often blurred, not immediately recognizable, and masked through humor or seemingly harmless intentions [144]. While microaggressions can occur in both on- and offline settings, our work focuses on racial microaggressions in online discussion communities. In recent years, it has been reported that an increasing number of Black individuals feel neither safe nor comfortable discussing race-related issues or experiences of racism and other potentially sensitive topics on social media [40, 63]. Despite the vocal prominence of public figures and influencers who are increasingly becoming more visible at the forefront of race-related conversations on the internet [98], about 43% of regular Black users report feeling anxious when it comes to discussing race-related matters publicly online [40]. Fearing risks of harassment, hate speech, and invalidation of lived experiences [64], some even choose to self-censor their views on racism or keep personal experiences of racism strictly private, especially on social media [8, 9]. Further, it is not just backlash from others, but also content moderation and hate speech policies by social media companies that seem to stifle conversations about race and racism for Black users [63, 64, 80]. For instance, Facebook and Nextdoor have been reported to remove posts shared by Black users regarding their personal experiences with racism either through automated filtering or through moderation [4, 43, 141]. Some Black users have reported being banned from the platform altogether [53] or locked out of their accounts for several hours or even days, a punishment known as being sent to "Facebook Jail" for sharing their views or experiences on race-related topics [63].

Human-Computer Interaction (HCI) scholarship in content moderation has shown that both human moderators and moderation systems disproportionately and often erroneously chastise users who often post about issues related to their marginalized identities, leading to false positives in content moderation decisions [38, 64]. At the same time, moderation systems often fail to detect race- or gender-based microaggressions that target marginalized users [61], allowing harmful content to remain online as false negatives [23, 104]. This is understandable given the significant topical and linguistic overlap between false positives and negatives, which can make it challenging for content moderation systems and human moderators to distinguish between the two. Such challenge underlies the premise of our work. In this work, we motivate the need to examine both false positives and false negatives in tandem, specifically in the context of acts and recalls of racial microaggressions. Through this work, we examine the linguistic signature of online racial microaggressions (acts) and how it differs from that of personal narratives recalling experiences of racial aggressions (recalls) shared by Black social media users, by asking the following questions:

- **RQ1:** *How can we leverage state-of-the-art language models to differentiate acts and recalls?*
- **RQ2:** *What are the similarities and differences between acts and recalls in terms of:*

  a) Themes: *what themes best characterize acts vs. recalls?*

  b) Contexts: *what contextual words are most predictive of acts vs. recalls?*
- **RQ3:** *What are the similarities and differences in the linguistic signature of acts and recalls?*

We answer these questions in the context of racism in the United States from the perspective of Black individuals. Through this work, we present a manually curated corpus of acts (2000 posts) and recalls (1264 posts), which were hand-annotated and iteratively verified through a workshop with Black participants. We use techniques in natural language processing (NLP) along with in-depth qualitative analyses to examine the language underlying acts and recalls of racial microaggressions with an aim to comparatively understand the lexical patterns that differentiate the two. Our findings show that state-of-the-art neural classifiers are able to distinguish acts and recalls with relatively high accuracy (95.4%) while more traditional language models based on n-grams features do so less efficiently (RQ1). While acts and recalls are thematically, contextually (RQ2), and linguistically (RQ3) distinct, they also share certain themes (appearance, criminality, ability, personality, and sexual exoticism) and key linguistic signatures (use of first person pronouns and out-group language) that can make it challenging for platforms and human moderators to distinguish between acts and recalls. Our findings represent an initial step towards better understanding semantic differences between acts and recalls of racial microaggressions on social media platforms and a re-evaluation of whether and how current socio-technical systems are able to differentiate between false positives and false negatives. Together, we aim to understand how Computer-Supported Cooperative Work (CSCW) research can best support members of marginalized groups. We argue that it is crucial to differentiate between acts (false negatives) and recalls (false positives) and utilize this knowledge to build and enhance online systems. By doing so, we can create constructive and safe online spaces where users can discuss, share, and learn from dialogues on race and racism with others. This understanding is vital in enacting inclusive and supportive online environments. Our contributions are as follows:

- We provide an in-depth characterization of the themes and linguistic signatures that underlie acts and recalls of racial microaggressions on social media communities, which has not been empirically examined at-scale by prior research to the best of our knowledge. Our insights show both distinct and shared themes and language patterns across acts and recalls, highlighting key challenges faced by content moderation systems and human moderators in their effort to distinguish between false and negative positives.
- Unlike the publicly available, generic off-the-shelf toxicity detectors that merely provide a score without an underlying explanation, in our comparative analysis of acts and recalls, we go beyond just classifying between the two, by leveraging interpretation techniques in deep learning (DL) to explain our classification results. Specifically, we address the lack of interpretive insight, typically associated with large pretrained language models by using Integrated Gradients [130] to identify contextual words that are most predictive of acts and recalls of racial microaggressions associated with Black users. By doing so, we overcome the black-box nature of DL language models by

making our classification results explainable – a practice we believe is contextually crucial when working with textual corpora such as ours that contain semantically nuanced and subjective content.

- We complement recent efforts in the NLP community to capture and surface implicit hate speech and microaggressions online [19, 46], by providing a dataset that is more specific to the context of race and racism in the U.S. Our data is hand-annotated and validated by Black participants whose labels we used as gold-standard truths to resolve any discrepancies between non-Black annotators. We make our dataset publicly available for the wider research community in hopes that it would serve as a benchmark in examining language associated with racial biases and microaggressions.
- We provide insights from our workshop discussions with Black participants to further inform and validate our findings. Given their racial identity as Black individuals and the contextual familiarity with the content of our data, we corroborate our findings with the rich insights from our participants. By doing so, we aim to strive towards the goal of directly incorporating the experiences and views of the marginalized in CSCW and HCI research.

Chapter 2

Review of Literature

2.1 Racial Microaggressions and Implicit Biases

Microaggressions are often subconscious [50, 127] or even unintentional [96, 126, 144], meaning that they are driven by an individual's implicit biases toward people who are not members of one's own in-group [86, 123]. In the context of racial microaggressions, social psychologists argue that most people generally do not deliberately exhibit or act on racial biases all the time [70, 87, 106]. Instead, most racial biases today often take implicit forms and manifest through social conditions to which people are exposed to, or through which they interact with others [106]. Nonetheless, implicit racial biases can have consequential damages. For example, researchers found that teachers were more likely to discipline Black students on their first rather than second offense, implying that instructors were quicker to see so-called "patterns" of bad behavior in Black children compared to those of other races [102, 116]. In another study, people who saw images of Black families tended to associate them with poorer and less safe neighborhoods, despite how middle-class those families appeared in the pictures [16]. Even when people generally do not consciously harbor racist views, research has shown that they tend to subconsciously link criminality and primitiveness with Blackness [20, 59, 109, 121]. In essence, people's cognition can subconsciously interact with the conditions they are exposed to when determining responses to other people, especially in the context of race [116]. Such a premise suggests that social conditions shape the nature of interactions, and this is not exclusive to offline realms.

2.2 Racial Microaggressions in Offline vs. Online Settings

While racial microaggressions that occur face-to-face may be difficult to respond to, people can still immediately call out the transgression as it presently occurs (or has just happened), especially when the offender is visibly identified and present. In fact, there are a countless number of guidelines providing recommendations on how to respond to racial microaggressions in various contexts [47, 90, 136]. However, such guidelines are specifically tailored to in-person, offline interactions, which occur in settings much different from those of online environments. According to [127], an example of an offline racial microaggression would be a White person checking their wallet or clutching their bag as a Black man approaches or passes them, insinuating a sense of fear that Black people are most likely to be criminals. Online, racial microaggressions often surface on social media as posts and comments reacting to content posted by Black users (e.g., "You're too pretty for a Black girl").

The different affordances of on- vs. online settings in which racial microaggression occur can potentially shape how victims experience or react to such transgressions. For example, online, offenders can hide behind the anonymity afforded by throwaway accounts, which can sometimes disinhibit bad behavior [5]. While de-identified settings are necessary and crucial in circumstances where users disclose sensitive personal experiences [5] or exchange support in stigmatized contexts [114], the affordance of online anonymity can make hate speech and harassment effortless [7, 99]. Further, throughout networked publics on social media, single users can easily connect to a congregation of thousands of others, meaning the scale of interaction and exposure can be one-to-many [1, 18]. While there are advantages to such scalability [43, 81, 139], this also means that users sharing or recalling personal experiences of racism online, can potentially face an army of aggressors who can instantaneously pile on their post by flagging or downvoting content or trolling in the comment threads. This can lead to an uptick in engagement metrics that may trigger content moderation systems to automatically flag the post for further review or removal for potentially violating platform policies [38]. Hence, scaled interactions in online settings can make it difficult for users to share recalls of racial microaggressions, or even personally respond to acts without the burden of risking oneself against mob harassment [100, 119]. Furthermore, unlike racial microaggressions that occur in person, online acts of racial microaggressions are rarely directed at individuals. Instead, the stereotyping language of acts often targets racial groups as a whole in generalized expressions [37, 38, 148], making it harder for users to directly call out the aggressor's offense based on personal grounds beyond the context of one's race.

2.3 Language as a Condition of Online Discourse

Scholars in CSCW and HCI studying online discourse have shown early on that the conditions through which social interactions take place inevitably shape the nature of such interactions [112, 113, 115]. It is therefore not just the design of platform features or algorithmic content-ranking, but also the language users regularly encounter through others that characterize the conditions of how people come across and talk about certain topics [113]. For example, when people process contentious issues online, negatively charged affective words induce more negative conclusions in ensuing discussions [115]. As such, linguistic patterns can effectuate negativity biases toward the subject of discourse among users [11, 22, 93, 125, 138]; language that characterizes online discussions on race or racism is not an exception [112, 113]. Further, while it is important to recognize that racism expressed through language can and does take extreme and overt forms, our present study focuses primarily on the more subtle and implicit expressions of racial biases in the context of online racial microaggressions targeting Black users. We do so for several reasons. First, research has shown that racial microaggressions can profoundly impact people's physical and mental health, self-esteem, and academic performance [14, 65, 78, 96]. However, the implicit nature of racial microaggressions can be camouflaged across everyday life [126, 144], such that the impact of offense and harm is often unrecognizable or perceived as insignificant [137], especially by the offenders (and bystanders), while receivers are relegated to self-doubt and distress [2, 128]. Second, while automated detection of explicit racial profanity (e.g., via customized lexicons or regular expressions) is currently possible and widely implemented across online platforms, detecting the more nuanced and inconspicuous language around racial microaggressions masked through everyday language is not yet systematically feasible at-scale [86, 125]. Finally, human content-moderation too can be predisposed to the moderator's own unconscious biases and subjective understandings of racism and race-related matters, which makes drawing the line between acts and recalls of racial microaggressions difficult. As a result, racially marginalized users repeatedly face the burden of navigating and resolving situations where they are censored, locked-out, or banned from their accounts for sharing personal views or experiences of racism or race-related issues [63]. Such experiences aggregated over time can invalidate or elicit self-doubt towards the user's lived experience as a racial minority [6, 66]. Such issues are precisely the challenges we bring attention to and aim to address through this work.

2.4 Understanding False Positives and False Negatives in Tandem in Content Moderation Decision-Making

Content moderation research in CSCW and HCI has shown that people from marginalized communities are disproportionately affected compared to other users [38, 52, 64?]. For example, trans and Black users are more likely to become victims of content moderation false positives, meaning that their comments are mis-classified and censored as harmful even when they do not violate platform policies [64]. Likewise, false negatives (toxic content that violate platform policies, but are left undetected) and their impact on users has been well documented in prior HCI literature [23, 103, 110]. For example, [104] shows that content moderators removed only one in 20 comments violating macro Reddit moderation norms in 2016, and one in 10 violating comments in 2020, highlighting that some categories of violation were more likely to slip through the cracks, leaving most anti-social behaviors unmoderated. This is because false negative content, which often includes microaggressions or implicit hate speech, are rarely explicit [47, 48, 68] and tend to be subtly disguised through humor [49], insider expressions, and neologisms [17]. As a result, this makes it harder for language models, let alone even human moderators, to identify false negatives [21].

Although content moderation researchers in CSCW and HCI have examined both false positives [38, 64, 73] and negatives [104, 110], as well as the impact they have on users [52, 57], such studies have generally examined the two discretely rather than in tandem. Our research motivates the need to examine false positives and negatives in conjunction. The significant topical and linguistic overlap between recalls and acts of racial microaggressions [64?] makes it difficult for content moderation systems and human moderators to detect the two apart [38, 147], often leading the former to be censored as a false positive and the latter to remain on the platform as a false negative [64, 83]. Incorrect content flagging [25, 132], or the inability of moderation systems and human moderators to differentiate when someone is critiquing racism (false positive) versus being racist (false negative) [26, 74], can be a gate-keeping practice that can evolve into a form of digital gentrification [52?], further exacerbating disparities between platform members and content moderators [117]. Our work aims to address such challenges highlighted by prior CSCW work in content moderation by examining the intertwined relationship between false positives and false negatives, specifically in the context of race-related social media posts. In so doing, this work takes a step towards building content moderation practices that aims to distinguish between false positives (recalls of racial microaggressions) and false negatives (acts of racial microaggressions) on social media.

Furthermore, most commercial toxicity models (e.g., Perspective API) used in social media content moderation systems, are not designed to distinguish false positives and false negatives in the first place, but to merely assign toxicity scores. As a result, content moderation algorithms that rely on these toxicity scores have difficulty differentiating online acts of racial microaggressions (false negatives) and discussions recalling experiences of racial microaggressions (false positives), often penalizing users they are supposed to protect [58, 72]. We explicitly address this gap by training a language model to comparatively learn the linguistic and topical nuances that are intertwined between false positives and the false negatives.

2.5 Lack of Explanations in Content Removal Decisions on Social Media

One major concern highlighted by prior research examining marginalized user's experiences with online content moderation is that most moderation systems fail to explain content removal decisions [74, 147]. Users who frequently experience content removals are often given vague (e.g., violated terms of service) or no explanations at all [64, 86]. As a result, they are left with very little knowledge as to what part of their language may have caused their post to be removed in the first place [143, 146]. Such lack of transparency and context around content removals [76, 108] make users feel frustrated and helpless as they are left unsure of how to subsequently engage on the platform [143]. Prior research has argued that enhancing the explainability of moderation decisions can not only empower users [45, 129], but also moderators as well [124, 131]. Moderators who rely on semi-automated crowdsourcing [55, 67, 88] or AI-led moderation [82, 83] experience less cognitive burden, stress, and reduced symptoms of Post Traumatic Stress Disorder (PTSD) [29, 32, 122] compared to those who do not [124]. However, the semantic and topical overlap between recalls and acts of racial microaggressions can make it difficult for content moderation systems and human moderators to detect the two apart, leading to false positives and negatives. Decision-making around grey content areas [64, 140], such as microaggressions, are also heavily subject to the moderator's personal and often limited understanding of what is and is not microaggressions [64], which can often lead to arbitrary and inconsistent removal justifications [134].

We address this challenge by providing a computational approach that makes classification decisions around false positives (recalls) and false negatives (acts) interpretable. In our work, we not only build a model that can distinguish between false negatives (acts) and false positives (recalls), but also provide a computational approach that helps contextualize as to what contributed most to the model's classifications decisions. Such insight can not only help moderators learn and understand the nuanced differences between acts and recalls, but also inform their decision-making as well. In our work, we use a deep learning interpretation technique to demonstrate which input feature (word) of a given post contributes most to the output decision as to why that post is most likely to be a false positive or a false negative. In so doing, we provide a scaled approach in help contextualizing and explaining model decisions in classifying between acts and recalls.

2.6 Challenges in Detecting Toxicity Through Language Models

CSCW and HCI scholars have demonstrated early on, both the effectiveness and limitations of prior approaches in identifying and curbing online toxicity, such as crowdsourcing [75, 85**?**], nudging user behavior (e.g turning off comments) [3, 120], and human moderation [24, 83, 145]. However, the challenge of countering the growing volume of toxic language has yet to be resolved [91, 149]. This is perhaps due to the diversification of toxic language [13] and online hate speech that increasingly contain neologisms [94], coded expressions, or subtle and indirect phases that mask harmful language [62]. As a result, detection of linguistic toxicity is difficult for both humans and machines [135, 150]. Recent scholarship in NLP has aimed to uncover such linguistic patterns by utilizing neural models to detect implicit hate speech [46, 151] and microaggressions [19, 68]. However, such studies use data that focus on a broad array of topics. Our work adds to this existing body of research by providing and analyzing a dataset that is more contextually targeted to the topic of race and racism in the U.S.

Furthermore, a growing body of NLP research in language and fairness has highlighted how language models unintentionally capture, reflect, or even amplify various social biases that manifest in the data they are trained on [15, 118]. For example, linguistic models that power YouTube's automatically generated captions tend to identify the language spoken by male and white users more efficiently than they do of female and minority users [36]. Scholars have also demonstrated racial disparities in NLP systems by showing how widely used commercial, off-the-shelf language models fail to recognize African-American English compared to other dialects [12]. Furthermore, state-of-the-art language models pretrained on large amounts of data from the internet collected at specific points in time are susceptible to learning unintended biases towards real-world entities [107]. For example, even though the phrases, "I hate Justin Timberlake" and "I hate Rihanna" both express the same semantics based on identical constructions, language models tend to classify the former as significantly more toxic than the latter [107], exhibiting gender disparity in toxicity scoring. Such issues may arise from disagreements in label annotations, especially when it comes to annotating texts associated with gender and race-related content – a task that is immensely difficult for human annotators to reach consensus around what is considered ground truth [42, 95]. Throughout our analyses, we strive to be aware of such aforementioned biases. Hence, we host a workshop with Black participants through which we iteratively discuss, learn and validate our labels of acts, and use the annotations from the Black participants as the gold standard to resolve any discrepancies between non-Black annotators.

Finally, another problem with neural classifiers is that the results from such models are difficult to understand due to their lack of interpretability [51, 77]. Therefore, users across online platforms that rely on toxicity classifiers powered by neural networks might question how their posts were evaluated [74, 146]. Therefore, as a step towards encouraging linguistic tools that can provide end-users an explanation as to why their post was (mis)classified as toxic, we interpret the classification result from our best performing language model that predicts acts apart from recalls of online microaggressions.

Chapter 3

Methods

3.1 Data Collection

Type	List of subreddits
Acts	askblackpeople, askScience, BlackPeopleTwitter, casualconversation, circlejerk, confessions, darkjokes, darkjokeunlocked, explainlikeimfive, Forwardsfromgrandma, gatekeeping, hiphopcirclejerk, insanepeoplefacebook, Jokes, offensivejokes, outoftheloop, pewdiepiesubmissions, relationship_advice, shitliberalssays, shitredditsays, shitaskscience, shittylifeprotips, showerthoughts, subredditdrama, Tankiejerk, TrueOffMychest, TrueUnpopularopinion, unpopularOpinion, WhitePeopleTwitter
Recalls	Blackladies, Cptsd_bipoc, datingadvice, interracialdating, Mixedrace, TwoXChromosomes
Both	askReddit, nostupidquestions, Teenagers, TooAfraidToAsk

Table 3.1: List of Subreddits Used to Retrieve the Posts and Comments for Acts and Recalls.

We introduce a new corpus called Recalls and Acts of Racial Microaggressions (RAMA) containing 2,000 instances of acts and 1,264 recalls of racial microaggressions from posts and comments from Reddit and Tumblr. Our data consists of acts and recalls of racial microaggressions, specifically against Black people. For Reddit, two authors first manually examined acts and recalls across posts and comments on subreddits known to contain

Type	List of search keywords
Acts	‘African American people’, ‘African American men’, ‘African American women’, ‘African American individual’, ‘African American individuals’, ‘African American person’, ‘African American man’, ‘African American woman’, ‘African American girl’, ‘African American boy’, ’African American ladies’, ‘African American guy’, ‘African American gal’, ‘African American lady’, ‘African American dude’, ‘African American kid’, African American chick’, ‘African American parent’, ‘African American student’, ‘black people’, ‘black boy’, ‘black girl’, ‘black men’, ‘black women’, ‘black man’, ‘black woman’, ‘black person’, ‘black individual’, ‘black individuals’, ‘black dude’, ‘black guy’, ‘black gal’, ‘black lady’, ‘black ladies’, ‘black chick’, ‘black kid’, ‘black parent’, ‘black student’
Recalls	“Microaggressions”; “Microaggressions, I face as a Black” + man, woman, kid, girl, gal, individual, person, guy, lady, dude, parent, student; “As a black” + man, woman, kid, girl, gal, individual, person, guy, lady, dude, parent, student; “I’m a black” + man, woman, kid, girl, gal, individual, person, guy, lady, dude, parent, student; “I’m an African American” + man, woman, kid, girl, gal, individual, person, guy, lady, dude, parent, student

Table 3.2: List of Search Terms Used to Collect Acts and Recalls From Reddit.

racial microaggressions (e.g., r/showerthoughts, r/TooAfriadToAsk, r/unpopularopinion)[1]. As subreddit profile pages display a list of other similar subreddits, we used this information and adopted a snowball approach to expand our list of subreddits to manually look for posts and comments containing acts and recalls of racial microaggressions targeting Black people. Upon examining approximately 150 subreddit pages, we finalized our list of subreddits to those shown in Table 3.1. Using these subreddits, three researchers then manually verified 300 acts and 200 recalls of racial microaggression posts and comments across a diverse array of topics. We then used these 500 posts and comments to identify an initial set of common keywords to be used in the API search as shown in Table 3.2. We iteratively expanded on the search keywords through multiple discussions to ensure their relevance and search strength. These keywords are similar to those used in prior work examining posts containing or recalling experiences of microaggressions [19]. We then used these keywords to collect posts and comments using Reddit’s official API PRAW [89] and pushshift.io[2]. A randomized selection of posts and comments collected from the API search were then inspected by the authors

[1] We identified that a large share of the acts in our data came from comments reacting to recalls on r/BlackPeopleTwitter.

[2] https://pushshift.io/

and annotated and verified by workshop participants (refer to Section 3.2).

Furthermore, we also added posts and comments from a Tumblr website[3] that contains a collection of self-reported accounts of acts of microaggressions across various topics [19]. We scraped all posts and comments topically pertaining to racial microaggressions and manually verified for acts and recalls related to Black people. Data from Tumblr were also annotated and verified by workshop participants.

3.2 Participant Workshop: Verifying Labels

In order to verify and annotate our RAMA corpus, we hosted workshop sessions with a total of 15 participants (6 Black and 9 non-Black, see Table 3.3 for demographic information). The purpose of the workshop was to obtain high-quality human labels as to whether or not a given social media post or comment contained an act of racial microaggression against a Black person/people.

We sent out workshop flyers through campus mailing lists and contacted respondents through email. Participants were then invited to a 90 minute workshop session, which was held over lunch that was provided and paid for. We conducted a total of two separate, but procedurally identical sessions - one with Black and another with non-Black participants. We used the labels from the Black participants as the gold standard to resolve any discrepancies between non-Black annotators. We began each workshop by introducing ourselves, the project background, key research motivations behind identifying acts vs. recalls of racial microaggressions, and the take-home annotation task. We aimed to minimize participant fatigue and encourage a safe discussion environment given the difficult nature of the topic and sensitive content associated with racial microaggressions. Hence, participants were informed that they could take a break or leave at any point during the workshop and that there were

[3] https://www.microaggressions.com/

no time constraints for the take-home task. In addition, participants had the opportunity to receive research credits by obtaining approval from their respective course instructors. This meant that their active participation in the workshop could be recognized and counted towards fulfilling their research credit requirements.

Once participants introduced themselves to one another, we used a guided PowerPoint presentation to explain and walk through multiple examples of comments and posts that contained acts and recalls of racial microaggressions across various themes (Table A.1), and invited participants to reflect on these examples and to share their perspectives.

We then proceeded to annotate 20 examples as a group through discussions. During the discussion, participants shared why they did or did not choose to classify a post/comment as an act of racial microaggression. Participants explained what part of the comment or the post/comment specifically contributed to their decision. At the end of the workshop, we invited participants to annotate 300 random samples as a take-home task. There was no requirement to further participate nor a deadline imposed on the take-home tasks. All workshop material (presentation guidelines and annotation examples) were shared with the participants at the conclusion of the workshop.

All workshop sessions were audio-recorded for transcription with participants' consent. Once we received the completed take-home tasks from participants, we checked for inter-annotator agreement. Labels from Black participants were used as ground truth values to resolve discrepancies among annotations from non-Black participants. Annotation agreement across all acts based on Fleiss squared kappa was substantial (k=0.77) [84].

ID	Racial-Ethnic group	Age	Gender	Ongoing/ Highest Degree
P1	Black/ African American	19	M	BA
P2	Black/ African American	20	M	BA
P3	Black/ African American	20	M	BA
P4	Black/ Caribbean American	20	F	BA
P5	Black/ Caribbean American	21	F	MS
P6	Black/ Caribbean American	25	M	PhD
P7	South East Asian	24	M	PhD
P8	South Asian	24	F	MS
P9	Middle Eastern	26	F	PhD
P10	Middle Eastern	25	F	PhD
P11	South Asian	25	F	MS
P12	South Asian	25	M	MS
P13	Asian American	26	F	MS
P14	South Asian	21	F	MS
P15	South Asian	25	F	BA

Table 3.3: Participant's Demographic Data.

3.3 Analysis

RQ1: How can we leverage state-of-the-art language models to differentiate acts and recalls?

To answer RQ1, we initially tested traditional ML-based classifiers known to work well with small amounts of data. To further understand the challenges of implicit hate detection and achieve compositional understanding beyond simple keyword-matching, we fine-tuned large language models (LLMs) such as BERT, RoBERTa and XLNet to distinguish acts and recalls of racial microaggressions. First, using a 80-20 split on the RAMA corpus, we balanced the distribution of the target class in our data in both train (80%, N=1680) and test (20%, N=420) sets. For the traditional ML models, we used the Naive Bayes (NB), Support Vector Machine (SVM) and Logistic Regression (LR) with standard unigrams, Term Frequency-Inverse Document Frequency (TF-IDF), and GloVe embedding [105] features. We used scikit-learn's *feature_extraction* attribute to extract features and *CountVectorizer* module with a (default) minimum word frequency $= 2, (2, 2)$ n-gram range. We used k-fold cross validation with $k = 5$ to avoid overfitting of the models. For neural models, we fine-tuned BERT, RoBERTa and XLNet and set the batch size to 16 with 8 training epochs, and used AdamW for optimization with 2e-05 learning rate. All baseline ML models were implemented using sklearn[4] and LLMs using PyTorch[5].

RQ2: What are the similarities and differences between acts and recalls in terms of:

a) Themes: what themes best characterize acts vs. recalls?

b) Contexts: what contextual words are most predictive of acts vs. recalls?

While state-of-the-art neural models are effective at high-level hate speech classification, they are not effective at spelling out more fine-grained categories with detailed explanations of the implied message [147]. First, to address this, we employ Integrated Gradients (IG) [130], a model interpretability technique in deep learning (DL) that helps identify key input features that contribute most to the model's predictive decision [**?**] and are calculated by computing the gradient of the model's prediction output to its input features [54]. IG (1) provides intuitive explanation for output decisions from transformer based models like BERT that often lack interpretive insight [27]. We used the best performing classification model in RQ1 along with IG to derive words predictive of the model's decision (RQ2a). In IG, a feature's contribution to the output of a neural classifier is calculated by considering the gradient of the model prediction with respect to that of the input feature. Integrated Gradients calculate the average of gradients at all points along a straight line from the baseline $x^{'}$, which is set to zero vector for text-based models to input x [27]. Formally, if : $R_n \rightarrow [0, 1]$ represents BERT, then the integrated gradient of the i-th dimension is:

$$\text{Integrated Gradients}\,(x; F) = (x - x_i')\, x \int_{\Theta=0}^{1} \partial F\,(x' + \Theta\,(x - x')) \div \partial x' d\Theta \tag{3.1}$$

[4]https://scikit-learn.org/

[5]https://pytorch.org/

The final attribution score of a particular word is the sum of the integrated gradients for each dimension of that word's embedded vector [39]. Second, we qualitatively examined and analyzed acts and recalls containing the top 50 IG tokens pertaining to each class from RQ2a and grouped them into most frequently occurring contexts. Finally, we created thematic categories to which these tokens belonged, based on the highest frequency of contextual use in the data (RQ2b).

RQ3: What are the similarities and differences between the linguistic signature of acts and recalls?

To analyze the rhetorical styles of acts and recalls, we employed an iterative open coding procedure. Initially, we approached the transcribed text through axial [30] and thematic coding [56]. Subsequently, to gain a deeper understanding of how language was used to convey underlying intent, we conducted Critical Discourse Analysis (CDA) [54]. A brief explanation of CDA will follow later in this paragraph. Two of the authors independently coded a test sample of 100 randomly selected acts and recalls drawn from the larger racial microaggressions dataset, then discussed each post together with assigned codes to establish a shared set of categories for acts and recalls respectively. The same process was then applied to another 100 randomly selected recalls, with detailed discussions held for each one. Following the creation of the resulting codebook, the authors proceeded to code a final set of 100 randomly selected acts and recalls. In our examination of linguistic patterns across acts and recalls, we utilized CDA as a methodological framework [54]. CDA offers valuable insights into how individuals employ language in different contexts to convey their underlying intent. This approach has been widely employed in CSCW and HCI scholarship, particularly in the analysis of social media comments [44, 113, 115]. By employing CDA, we aimed to identify similarities and differences in linguistic patterns between acts and recalls, focusing on how posters conveyed their underlying intent. To enhance the validity of our qualitative findings, we also engaged in workshop discussions with Black participants, thus validating and enriching our analysis.

Chapter 4

Findings

In this chapter, we provide our findings related to classifier performance for differentiating acts and recalls of racial microaggressions (RQ1). Our findings revealed higher level differences between acts and recalls across various themes identified via Integrated Gradients (RQ2), as well as more granular differences in their linguistic signatures via Critical Discourse Analysis (CDA) and focus group interactions (RQ3).

4.1 RQ1: Classifying Acts vs. Recalls

To answer RQ1, we first built natural language based classification models to classify acts vs. recalls of racial microaggressions. We experimented with several traditional models such as Support Vector Machine (SVM), Naive Bayes, and Logistic Regression as baseline classifiers along with three different feature extraction methods (GloVe, TF-IDF, n-grams), as well as start-of-the-art neural models such as XLNet, BERT, and RoBERTa, to classify a user post/comment into an act or recall. We observed that RoBERTa and XLNet achieved a high level of performance, with accuracies at 0.954 and 0.917, respectively; in contrast, SVM and BERT had far lower accuracy scores at 0.760 and 0.906, respectively. The modeling results

	Binary Classification Result			
Model	Precision	Recall	F-1	Accuracy
SVM (n-grams)	0.752	0.747	0.749	0.716
SVM (TF-IDF)	0.639	0.653	0.645	0.618
SVM (GloVe)	0.709	0.735	0.721	0.708
Naive Bayes (n-grams)	0.717	0.721	0.718	0.722
Naive Bayes (TF-IDF)	0.622	0.608	0.614	0.621
Naive Bayes (GloVe)	0.667	0.525	0.587	0.545
Log. Regression (n-grams)	0.633	0.692	0.661	0.693
Log. Regression (TF-IDF)	0.769	0.724	0.746	0.719
Log. Regression (GloVe)	0.719	0.72	0.719	0.733
BERT	0.864	0.925	0.893	0.906
XLNET	0.891	0.968	0.927	0.917
RoBERTa	**0.921**	**0.950**	**0.934**	**0.954**

Table 4.1: Classification Metrics for Acts vs. Recalls (Best Performance Is Bolded)

are presented in Table 4.1.

4.2 RQ2: Interpreting Acts vs. Recalls

To explain the predictive decisions of our best-performing neural language classifier, we leverage Integrated Gradients (IG), a model interpretability technique, to extract key words that the model considered most predictive of acts vs. recalls. We use IG to identify specific tokens that contribute most to the model's predictive decision by computing attribution scores for each of the tokens. Attribution scores indicate how much a specific token correlates to the model's prediction judgment. We categorized the top 40 tokens with the highest and lowest attribution scores ranging from positive (predictive of recalls) to negative (predictive of acts) values by each class (act/recall), with the magnitude indicating the predictive strength for each class. Table 4.2 details the themes that emerged in the acts and recalls respectively. In addition to the attribution scores, we provide the number of times each word occurred in the corpus (Freq.). Tokens are listed in the descending order of their predictive strength and grouped by salient themes based on the most common contexts in which these tokens appeared across sentences. To understand how predictive words generated by IG are associated with themes that emerged across acts and recalls, we analyzed how these tokens were contextually used in sentences across posts and comments. The authors qualitatively analyzed the semantic contexts of how each word was used and categorized the tokens into salient thematic categories as shown in Table 4.2. Our analysis revealed three themes unique to acts, four themes associated with recalls, and five overlapping themes.

4.2.1 Themes of Acts

Questions Among posters of acts of racial microaggression, the tokens, *why* and *question,* are often used to pose a question along the lines of why a certain stereotype exists. Consider this example of an act: *"Why do black men have such dry hands?...I have just realized that while making this question that I've never [shook] the hand of a black woman."* The poster asks a question (*'Why do black men have such dry hands?'*) and then proceeds to acknowledge that they posed a question. By questioning the validity of their own question while writing the post, the poster appears to be thinking out loud, thereby not filtering their thoughts.

Ethnicity The tokens *ethnic, indian, hispanic,* and *latino* are often used by posters of acts to compare Black people to other races, especially racial and ethnic minorities, such as Indian and Hispanic people. For example, the poster of the following act creates a hierarchy of intelligence based on race in order to highlight their belief that Black people are less intelligent than other people, using different racial majorities (*'caucasians'*) as well as racial and ethnic minorities (*'asians,', 'indians,' 'latinos/hispanics'*) to elucidate their point:

"Some say that you can order races on their intelligence with asians on top, indians and caucasians a little below, latinos/hispanics about 1/4 std. deviation below whites, and blacks

about 2/3-1 standard deviation below whites."

Evolution and Human Race Tokens such as *earth* and *evolution* are frequently used by posters of acts to discuss where Black people came from as well as what led them to develop their distinct appearance and abilities. In addition to using the token *earth* to imply a point of origin, the poster also uses the word to casually place Black people outside of the human race: *"If Adam and Eve are the first people in the Earth and they are white, why are there Black people?"* Moreover, the following act uses the word *evolution* as a way to justify why Black people are faster than other races by connecting a common stereotype to evolution: *"Black people are faster because of evolution."*

4.2.2 Themes of Recalls

Relationships Posters of recalls commonly use the word *friend* to describe the perpetrator of an act of racial microaggression or the person that experiences a racial microaggression alongside them. The poster of this recall uses the word *friend* to describe the person that they were with when they both experienced a racial microaggression: *"I'm Dominican-American and one day me and my friend who's Bengali went to the mall. We walked into a MAC store and a White lady approached us and asked us: Where are you guys from? You guys look exotic."* Our findings suggest that this *friend* is often times Black or another racial minority, such as Bengali. On the other hand, the word friend is also often used to describe the perpetrator of an act of racial microaggression: *"I had a white male friend of mine in high school tell me that I'm the darkest he would go."* Our findings also suggest that the token, *husband*, is often used to describe the person that the victim of the racial microaggression is compared to by the perpetrator of the act: *"After expressing his shock upon seeing my husband's last name (whose family is German) he said, "Oh, so you've got a good Jewish boy, huh?" You must feel lucky!"*

Workplace Recalls often contain tokens such as *job* and *interviews* to describe acts of racial microaggression that occur in the workplace: *"At job interviews, I tell them where I'm from, born and raised in the Dominican Republic, and they say "Oooh" with a tone of disappointment."* Here, the words *job* and *interviews* serve to specify where in particular the poster has experienced a racial microaggression.

Attributed to Acts			
Themes	**Token**	**Attribution Score**	**Freq.**
Questions	question	-0.077	29
	why	-0.169	218
Ethnicity	ethnic	-0.0147	9
	indian	-0.020	11
	hispanic	-0.029	6
	latino	-0.044	3
Evolution and Human Race	earth	-0.009	2
	existence	-0.017	5
	evolution	-0.017	2
	population	-0.017	6
	civilization	-0.038	3
	humans	-0.066	2
	primates	-0.015	2
Appearance	fat	-0.004	2
	monkeys	-0.006	2
	palms	-0.007	6
	gorilla	-0.009	4
	skinned	-0.024	28
Criminality	police	-0.003	6
	robbed	-0.014	2
	dangerous	-0.014	11
	commit	-0.016	14
	violent	-0.023	11
	streets	-0.056	5
	attacking	-0.058	3
Ability	sports	-0.003	7
	intelligent	-0.020	3
	IQ	-0.048	7
Personality	ignorance	-1.300	5
	scary	-0.001	3
	names	-0.004	15
	dumb	-0.004	5
	disgusting	-0.004	2
	funny	-0.007	6
	ghetto	-0.009	7
	loud	-0.018	11
Sexual Exoticism	sex	-0.017	10
	attracted	-0.037	12
	hotter	-0.005	7
	sexual	-0.065	4

Attributed to Recalls			
Themes	**Token**	**Attribution Score**	**Freq.**
Relationships	friend	0.1820	76
	husband	0.1597	12
	mother	0.1419	13
	partner	0.1004	7
	father	0.0987	14
	family	0.0848	37
	boyfriend	0.0697	17
Workplace	job	0.1751	5
	manager	0.1665	31
	company	0.1596	6
	office	0.1508	75
	teacher	0.1221	13
	interview	0.1122	14
	employment	0.104	54
	internship	0.1025	19
Everyday Life	cafe	0.1577	1
	salon	0.1464	1
	shopping	0.1340	3
	church	0.1195	4
	grocery	0.0941	4
	gym	0.0776	2
Geographical Location	california	0.2480	1
	london	0.153584	1
	midwest	0.2480	4
	europe	0.1228	3
	chicago	0.1185	2
Appearance	hair	0.0843	143
	straightened	0.0113	6
	ape	0.0003	2
Criminality	criminals	0.0078	3
	neighborhood	0.0017	9
Ability	basketball	0.0040	4
	athletic	0.0283	4
Personality	smelled	0.0115	2
	lazy	0.0110	3
	stronger	0.0034	4
	ignorant	0.0029	6
	stupid	0.0004	3
Sexual Exoticism	exotic	0.0466	6
	attractive	0.0322	12

Table 4.2: Themes Emerged From Attributive Words of Acts and Recalls With Scores in Descending Order.

Everyday Life Recalls often contain tokens relating to everyday activities or places that people frequent on a day to day basis such as a *café* in order to convey what they were doing when experiencing an act of racial microaggression: *"On Friday morning, as I walked to the café between classes at my predominantly white university, the school appointed photographer offered me a free coffee if I agreed to play the role of the cheerful token black woman in a group of strangers."*

Geographical Location Posters of recalls often use tokens relating to geographical location (countries, cities, regions, etc.), such as *California*, to communicate where they experienced a racial microaggression: *"I walk into a gas station market in California with about ten of my Latina/o and black high school students to buy snacks for our college road trip, and within five minutes, we hear, "SECURITY CHECK ON ALL AISLES."*

Overlapping Themes Across Acts and Recalls

Appearance Both acts and recalls contain words such as *gorilla* and *ape*, respectively. While posters of acts commonly use the word gorilla in order to imply that Black people do not deserve to be treated as human beings: *"Black people appear more closely related to gorillas than human,"* posters of recalls commonly use the word *ape* in order to describe instances in which they have been compared to such an animal: *"He was also the worst of the people making these jokes in high-school and shortly after, making hundreds of "all black people are criminals" jokes and *comparing me to an ape* one time."* As it relates to the topic of appearance, given that ape is an umbrella term that includes several species, one of which is a gorilla, it follows that acts within this theme are often more specific than recalls. Similarly, while acts utilize a variety of words such as *fat, palms, and skinned* with the purpose of negatively stereotyping Black peoples' appearances: *"Black people are fat because McDonalds is all they can afford"*, recalls frequently use words relating to a single feature, hair (*hair, straightened*) to describe their experience of being the victim of microaggressions about the texture of their hair: *"This is why I don't ever straighten my hair anymore,* *even though it's something I used to like to do for fun on occasion, because the compliments always seem to insinuate that my normal hair is unprofessional, unruly, or otherwise socially unacceptable."*

Criminality Posters of acts often use anecdotal evidence and a variety of words related to *criminality*, such as *police, robbed, and dangerous* in order to make stereotypical statements about Black people: *"Today I almost I got my car robbed from me. I've gotten robbed twice by a black person and this is the 3rd time but this time I was able to get away."* In contrast, posters of recalls frequently use a single word, *criminals*, in order to describe being associated with criminals: *"He was also the worst of the people making these jokes in high-school and shortly after, making hundreds of "all black people are criminals" jokes and *comparing me to an ape* one time."*

Ability While acts commonly contain tokens such as *sports* and *IQ*, recalls frequently contain tokens related to sports, such as *basketball* and *athletic*. For example, in this act the poster uses the word sports to highlight that Black people are only good at rap and sports: *"Black people of Reddit, How does it feel to be inferior and only good at rap and sports?"*. In addition, the acronym *IQ* is used in order to demonstrate that Black people are naturally less intelligent than other races, as IQ is seen as more of an inherent intelligence as opposed to intelligence derived from hard work: *"Black people have a way lower IQ across the board compared to their white counterparts."* In contrast, recalls focus solely on sports: *"The normal stereotypical things like us liking fried chicken and watermelon, every black person knows how to dance (my brothers are proof that's a lie) we're all good at basketball (I'm proof that's a lie)."* Finally, similar to criminality and appearance, recalls within this theme contain a fewer variety of words, indicating that the content of acts is more dispersed as compared to recalls.

Personality Posters of both acts and recalls commonly make use of tokens related to negative personality characteristics such as *dumb* and *stupid*. While acts frequently use the word dumb to criticize Black people: *"They never know what I'm talking about because they are dumb,"* recalls frequently make use of the word stupid to describe how they are perceived by others: *"So that's why I'm not only likely to be a thief, I'm likely to be a stupid thief!"*

Sexual Exoticism Posters of acts frequently use the tokens, *attracted* and *sex*, to highlight their perception of Black people as highly sexually desirable just because of their race: *"I'm extremely sexually attracted to Black men and women."* In addition, this act uses the word sex to underscore that being Black is a condition that must be satisfied for them to agree to have sex with someone: *"I will only have sex with Black men."* In contrast, posters of recalls frequently use the word attractive to describe that other races typically find their race unattractive: *"Like when people tell you aren't black because you're attractive."*

4.2.3 RQ3: Characterizing Acts vs. Recalls

To further understand the nature of acts and recalls of racial microaggressions in Reddit posts and comments, we examined the linguistic attributes manifest in their content using Critical Discourse Analysis (CDA) to better understand the social purpose underlying the linguistic patterns observed in acts and recalls of racial microaggressions. Tables 4.3, 4.4, 4.5 represent the linguistic pattern that emerged from the analysis and examples for each linguistic pattern unique to acts and recalls respectively. Our findings suggest three linguistic patterns in both acts and recalls of racial microaggression from our data. We utilize linguistic analysis as well as data gathered from our workshop participants to better understand the functional purposes of the similarities and differences we observed in the linguistic patterns of acts and recalls.

Linguistic Signature of Acts

Questions Our findings revealed questions to be a key linguistic pattern in acts of racial microaggression. Consider this example of an act: *"Why are black people so athletic? Not to sound racist, but I have recently noticed that black students excel at all the sports in my school. I am White and most people in my school are white, However the black players are the best in every single sport for our school. Why is that? I don't mean to be racist or offend anyone. Just wondering."* This statement serves to create a broad generalization of all Black people as being athletic, falling under several themes of racial microaggressions such as racial categorization and sameness, assumptions about intelligence, competence, or status, and connecting via stereotypes [144]. Consider the statement below from one of our workshop participants, P4:

> *"It is still a microaggression in the form of a question. You are still generalizing black people and you kind of believe that statement which is why you are curious about it" —P4*

This statement is a racial microaggression masked in the form of a question, which can give the statement more of a tone of curiosity than aggressiveness (refer to Table 4.4). Nevertheless, just like P4 highlighted, this curiosity is still a form of generalization. The poster of this statement seems to believe that all Black people are athletic, which is why they are curious about why that is the case. Consider the statement below from another one of our workshop participants, P5:

> *"There are only a few black players that are the best but when they are asking this question, they clearly generalize. If the star athletes weren't black, you wouldn't be wondering, "why are these people so athletic?" It would just be obvious that they probably practice a lot"-P5*

Despite the subtlety of this statement, P4 further highlights how this statement is a generalization, emphasizing that if the star players were white, the posters would likely attribute

Linguistic Pattern	Examples of acts of Racial Microaggression
Questions	• ***Why*** *are black people so bad at swimming?* • ***Why*** *do so many black people litter? I don't think a day goes by that I don't see someone throw trash out of their car, 9 times out of 10 that person is black.* ***Why****?* • ***Black people of Reddit****, which one of you stole my bike?* • ***Black people of Reddit****, can I touch your hair?*
Use of absolute terminology ('all', 'never', 'ever', 'should', 'absolutely', 'only')	• *[Black people] are* ***all*** *mentally handicapped and physically incapable of supporting themselves.* • *I would* ***never, ever*** *hire someone with a 'black' name on their resume. I wouldn't even interview them.*
Use of statistics	• *Black people are not oppressed and if they want to be in prison less, they should not be committing* ***53%*** *of all homicides while only being* ***12%*** *of the population.* • *12% of the population is black people, yet they commit so much more crimes.*
Use of modifying adverbs or adjectives ('most', 'usually', 'consistently')	• *Black people* ***usually*** *name their kids after stuff they can't afford. Like Mercedes, Diamond, Hope, and Insurance.* • *Black people are* ***consistently*** *the most rude, demanding, ignorant, of what want and shady.* • *I fail black students way more often because, objectively, they make the* ***most*** *mistakes on driving test.*

Table 4.3: Linguistic Patterns Observed Using Discourse Analysis for Acts With Examples.

their talent to practice and hard work instead of race. On the other hand, our findings also suggest that posters commonly disguise microaggressions using a combination of curiosity and humor. Consider this example of an act: *"Black people of Reddit, which one of you stole my bike?"* This statement uses a common stereotype of Black people, criminality, to make a joke in the form of a question. By directly addressing members of the Black community on Reddit *("Black people of Reddit")* and assuming one of these members stole their bike *("which one of you")*, the poster of this act utilizes humor to soften a harsh stereotype.

Use of Absolute terminology and statistics In addition to questions, our findings show that acts utilize absolute terminology and statistics in order to justify making a racial microaggression. Consider this example of an act: *"Black people are not oppressed and if they want to be in prison less they should not be committing 53% of all homicides while only being 12% of the population."* This statement uses the problematic "13/50" argument [79], which is commonly used to stereotype Black crime in order to make statements appear factual as opposed to stereotypical [69]. The 13/50 argument is an overused and often misleading talking point that poses that black people make up only 13% of the population but commit 50% of all known crimes [142]. Consider the statement below from one of our workshop participants, P3:

"There is a lot more to that statistic because the original intent of the police was to monitor Black people"-P3

According to P4, such statistical reference is biased because the United States history of systemic racism has resulted in Black people often being the target of police [10, 92]. Moreover, another participant highlights how the use of statistics matters in determining whether a statement should be considered a racial microaggression:

"There is a difference between using a statistic to prove that there is a problem with the

prison system versus using it to say something about Black people."-P4

In this statement, P4 highlights that she believes that using a statistic to make a statement about an institution, such as the prison system, is different from using a statistic to make a statement about a certain race, such as Black people. Given that institutions such as the prison system are not human, she implies that using a statistic to make a seemingly *"factual"* statement that is negative about a particular group of humans is bound to be hurtful to people. The poster of this act uses the statistic in order to justify making a racial microaggression on the grounds that they are unbiased and are just disseminating a fact.

Use of modifying adverbs or adjectives The last key linguistic pattern distinct to acts is the use of modifying adverbs/adjectives. Consider this example of an act: *"Black people usually name their kids after things they can't afford. Like Mercedes, Diamond, Hope, or Insurance."* In response to this statement, one of our workshop participants highlights why this statement is a stereotypical/generalized statement, stating that she doesn't know anyone with those names:

"This statement is too generalizing. I don't know a single person with those names. It is not as common as you think."-P5

The poster of this act uses the modifying adverb, *'usually'*, in order to characterize this behavior of Black people as frequent. Unlike the posters of acts that use statistics to justify making a racist remark, the poster of this act uses a colloquial term, *'usually'*. By using the modifying adverb, *'usually'*, the poster of this act seeks to give the impression that their personal knowledge is sufficient to justify such a claim. Thus, similar to posters of acts that utilize statistics, this poster tries to justify making a stereotypical racist remark.

Lingusitic Pattern	Examples of Recalls of Racial Micraoggression
‘White’ and ‘White people’	• *I wish **white people** in general would stop commenting on my appearance unless it’s to compliment me or to tell me that I have something stuck in my teeth.* • ***White people** singling me out at social events to make small talk with me about race/politics.*
Paste Tense Verbs	• *I **was** at work and the topics of racism came up with my boss who is Italian...* • *I **felt** irritated at having to explain that yes, I am a REAL programmer.*
’Only Black’	• *I had an English teacher who loved to talk, and whenever she’d say anything about race or Black culture, she’s turn to me(the **only black kid** in the room) as if to validate/confirm the statement.* • *I was the **only black girl** in the room with him and ten other coworkers*

Table 4.4: Linguistic Patterns Observed Using Discourse Analysis for Recalls With Examples.

Linguistic signature of Recalls

‘White’ and ‘White people’ One notable linguistic feature of recalls is the use of the phrases *‘White’* and *‘White people.’* Consider this example of a recall: *“White people singling me out at social events to make small talk with me about race/politics. I think they want to see me get impassioned or educate them. No I’m tired, I came out to have fun.”* In this recall, the poster is expressing his/her discomfort of being *‘singled out’* at social events by *‘white people’*. Clearly stating the subject *(‘white people)’* that is causing his/her discomfort, the poster uses this phrase to highlight the source of their tiredness. By making the source of their tiredness very clear, the poster seeks to have his/her experiences validated.

Past Tense Verbs Another common linguistic feature we noticed in recalls was the use of past tense verbs. Since recalls are recounts of acts of racial microaggressions, it follows that most recalls describe events that took place in the past. Consider this example of a recall: *"I was at work when the topic of racism came up with my boss who is Italian..."*. The poster of this recall is describing the setting of when he/she experienced an act of racial microaggression. The use of past tense verbs, such as *'was,'* and *'came'* aids in disclosing the setting in which the act took place. This is a key element in the poster's recount of their experience. Prior work in human communications research suggests that lying individuals use fewer words and fewer past tense verb forms [41]. These verbs help provide readers with a confidence that the poster must be telling the truth because they seem to be recounting their past experience very clearly. This type of disclosure helps create an overall tone of honesty and authenticity in the statement, as the poster hopes to have their experience validated by others.

'Only Black' The use of the phrase *'only black'* was the last prominent linguistic patter distinct to recalls that our findings revealed. Consider this example of a recall: *"I was the only black girl in the room with him and ten other coworkers."* The poster's use of the phrase *'only Black'* in contrast to *'him and ten other coworkers'* underscores her discomfort at being the only Black girl in the room.

Linguistic Similarities between Acts and Recalls

Use of First-Person One notable difference between acts and recalls of racial microaggression is the role that first person voice serves in context. Consider the example: *"I am not a racist, but it seems whenever I sit near a group of black people I can't hear the movie over all the noise they make."* Here, the poster uses first person to preemptively defend himself/herself from being called a racist. One of our workshop participants, P3, states that this defensiveness does nothing to absolve the poster of what he/she said:

> *"Just because you say I don't be mean to be such and such does not automatically resolve you of what you said before or after. Maybe they don't mean to be racist but words are*

words and the implication is still going to be there regardless of intent. "-P3

While this type of linguistic pattern appears to mask a racial microaggression, P1 points out that while the poster may not have bad intentions, the statement is still an act of racial microaggression, and therefore, the hedging does nothing to reduce the severity of the statement. Unlike posters of acts that seek anonymity, posters of recalls typically use first person voice to thoroughly describe themselves-*"I am a 22 year old, brown skinned African American girl. In school in Maryland. I felt out of place and isolated."* Here, the poster utilizes the first person voice twice to describe her age and feelings. Based on LIWC analyses [133] and manual inspection, our results indicate that roughly 40% of recalls that utilize first person do so to describe themselves. This self-disclosure serves to create an overall tone of authenticity and honesty, which the poster hopes will allow his/her experiences to be validated.

Moreover, our findings suggest that posters of acts utilize first person voice in even subtler ways to hedge the aggressiveness of their comments. Consider this example of an act from our dataset:

"Speaking as a capital "C" Conservative. I totally agree. I enjoy good, well developed characters, I don't care if they are black, gay, green, alien, or inanimate objects. Scandal is a great show, orange is the new black was great for the first couple seasons, and Jesus is a great character in the walking dead. (That's coming from an orthodox catholic conservative). Now to be honest, I hated black panther, just didn't think it was a good movie."

Workshop participants agreed that the last sentence of the post contains the microaggression; nevertheless, the poster makes several statements before in order to take attention away from it. The repetitive use of first person voice (*"I"*) prior to the last statement serves to highlight the poster's desire to portray themselves as an objective critic of entertainment. Similar to our previous findings, one of our workshop participants highlights that this superfluous build up is just another method of hedging a racist remark:

"They try to use color blindness to remove themselves from what they are saying"-P6

The poster seeks to show that they see everyone as equal by saying, *"I don't care if they are black, gay, green, alien, or inanimate objects."* By making a somewhat extreme statement, that they don't see color, the poster seeks to curb the aggressiveness of his/her statement. Moreover, one of our workshop participants, P3, highlights this type of hedging has consequences beyond just being a way to remove oneself from the consequences of making a racist remark:

"I definitely can take offensive to this since you are not validating who people are-you are not acknowledging where they came from or what they experienced"-P2

This type of false color blindness fails to acknowledge the systemic racism that has existed in American culture for many decades [144]. Nevertheless, like P3 points out, these types of statements are often invalidating for Black people if they are proud of their identity or have suffered because of it (for more details on the definition and examples of color blindness, refer to Table A.1, Appendix). While acts utilize superfluous build up and false color blindness to hedge the severity of microaggressions, our findings suggest that posters of recalls also incorporate build up prior to the point they are trying to make.

"I will ALWAYS be one to want to expand my viewpoints, appreciate history, and under the social climates, but I'm fucking tired... I 10000% care about race and gender issues within our communities, but I'm tired of including "others" in the conversation. How can we navigate things like institutionalized racism without begging "others" to see us as human? Unfortunately, it's never going to change. Maybe I'm just being a pessimist, but..."

Linguistic Pattern	Examples of acts of Racial Microaggression From Our Data	Examples of recalls of Racial Microaggression From Our Data
Use of First Person	• ***I'm*** *not a racist, but it seems whenever* ***I*** *see sit near a group of Black people,* ***I*** *can't hear the movie over all the noise they make.* • ***I*** *don't hate them,* ***I*** *don't bully them, but* ***I'm*** *careful with them, as if they were criminals.* • ***I know I'll be downvoted****, but anyone in the U.S. who works for tips knows that Black people are far less likely to tip...*	• ***I'm*** *an African American graduate student, and* ***I*** *teach at a large university.* • *It angers me when people measure my race by the way* ***I talk, dress and carry myself.***
"Us" vs. "Them" Language	• *Black redditors, what is your take on having a white friend? What do you see* ***us*** *as? Or any other race?* • *Maybe we shouldn't let* ***them*** *(Black people) vote.*	• *How can* ***we*** *navigate things like institutionalized racism without begging "others" to see* ***us*** *as human?* • ***They*** *always said it was a joke but* ***they*** *kept doing it over and over.*

Table 4.5: Linguistic Patterns Observed for Acts and Recalls Using Discourse Analysis With Examples.

By emphasizing the word *always* and using the first two sentences to prove that he/she is an open minded individual, the poster of this recall is trying to establish themselves as someone that is legitimate and trustworthy in the Black community. Therefore, the poster uses superfluous build up to gain peoples' trust and have his/her thoughts and feelings heard.

Use of "Us" vs. "Them" Language Another notable characteristic of acts and recalls of racial microaggression is the use of Us vs. Them language. Consider the example from our critical discourse analysis: *"Maybe we shouldn't let them (Black people) vote".* By utilizing both *we* and *them*, this statement is characteristic of Us vs. Them language. The juxtaposition of these words serves to highlight the presence of two different groups, *we* or *us* and *them* or *"Black people."* This creation of an in-group and out-group serves to portray Black people as second class citizens.

On the other hand, recalls utilize *Us* vs. *Them* language in order or to emphasize their feelings of being discriminated against. Consider this recall from our dataset: *'They always said it was a joke but they kept doing it over and over."* The use of the word *"they"* in this statement is characteristic of *"them"* language in *Us* vs. *Them* language. The sentence highlights that the out-group is hurting the in-group's feelings by making what they considered to be a joke about the victim's race. In context, the use of *Us* vs. *Them* language helps underscore the victim's discomfort due to the "joke" by highlighting the actions of the out-group.

Chapter 5

Discussion

5.0.1 Language Mimicry and Relational Dynamics in Online Discussion Communities

Online discussion communities, such as Reddit, naturally embody relational dynamics across users based on community roles (e.g., moderators/ admins vs. regular users) and membership statuses (e.g., old vs. new members). Research shows that such social structures and relational hierarchies within online discussion groups can potentially play a role in language mimicry and adoption. Language coordination is a phenomenon in which people tend to unconsciously mimic the language of others by responding with similar words or phrases [97]. Research in computational linguistics has demonstrated how language coordination persists across conversations in ways that reflect power differentials between people. For example, in Supreme Court case settings, lawyers tend to linguistically mimic the language of the Supreme Court justices rather than vice versa [31]. Such language coordination also occurs online: Wikipedeans tend to echo the linguistic style of admins significantly more than that of non-admins who are perceived to have a lower status within the community [31]. Our findings show that acts of racial microaggressions on social media embody persistent linguistic patterns, such as absolutist expressions (e.g., *never, ever hire someone with a Black name*) and modifying adverbs (e.g., *Black people are consistently the rudest*) that generalize or racially discriminate against Black people. Acts are also frequently masked in the form of questions disguised as genuine curiosity, or conveyed with statistics that tend to factualize selective information as broader truths. Given the presence of relational dynamics in online communities on top of platform affordances (e.g., up/down votes, likes, volume of comments) that interplay with such dynamics, linguistic patterns of acts can be mimicked and adopted across users, potentially amplifying racial biases, and endorsing harmful assumptions that underly acts of racial microaggressions. For future work, we intend to empirically capture how membership statuses and power dynamics within online discussion groups are associated with the adoption and spread of linguistic patterns of acts and recalls.

5.0.2 Critical Race Theory in Language and the Importance of Counter-Storytelling

According to Critical Race Theory, social conceptions of race and racism shape, and are shaped by laws, social movements, politics, and the media [101]. Such an argument is well-reflected across the theoretical premise of several anthropological research studies on race and language. Anthropological linguists have long recognized the importance of treating racial categories and concepts ”not as objective facts about the world, but as the outcome of discursive processes that operate across intersecting scales of space and time” [28]. That is because, the process through which language itself is racialized, or the way language racializes certain groups of people over time, inevitably involves linking certain objects, ideas, and themes to a racial group [60, 71, 111], thereby concretizing stereotypes about a particular race, as shown in our findings. For example, the dominant themes that emerge across acts tend to link Blackness with **crime, sexual exoticism, and questionable belonging to the human race**, which falsely perpetuate racial tropes about Black people's **personality** (e.g., *funny, loud, dumb, creepy, ghetto, etc.*), **ability** (*sports, intelligent, IQ, etc.*), and **appearance** (*fat, hair,* etc.). Interestingly, many of these identical themes appear in recalls, wherein Black users engage in what critical race scholars describe as counter-storytelling. Counter-storytelling is the act of recounting an individual's experience with racism, typically through language that operates as a discursive tool for challenging majoritarian perspectives in culturally dominant discourses on race [35]. Stereotypes perpetuated through racial attitudes, and conceptions of race and racism that have persisted across centuries, tend to become normalized into culturally dominant narratives [34]. As a result, implicit racism as observed through online acts of racial microaggressions in our data can falsely appear as race neutral. Black users, as shown in our findings, in essence, call-out such biases through counter-storytelling, through which they directly challenge racial stereotypes and attitudes by conversing on the same topics and themes present in the acts through autobiographical language. Sociotechnical systems that fail to distinguish acts and recalls risk suppressing these counter-stories shared by Black users. Both critical race scholars and historians argue that sharing personal stories has always been essential to the survival and liberation of racially oppressed groups [33]. Ensuring sociotechnical systems that safeguard rather than impede important conversations and experiential knowledge shared through counter-stories, such as the ones shown in this work, are critical to establishing more inclusive and enriching environments for online discourse.

Chapter 6

Conclusions

In our research, we urge for a deeper understanding of the subtle yet significant differences in the way racial microaggressions are expressed on social media. We believe it is crucial to re-evaluate how users and the current socio-technical systems discern between the acts of microaggressions and the recollection of those experiences. To initiate this endeavor, we gathered a collection of instances representing both the acts of microaggressions and the recollections of those encounters. These instances were discussed, carefully annotated, and validated by Black participants during a dedicated workshop session. Using this curated dataset, we conducted a comprehensive analysis to classify, interpret, and characterize the language used in both the acts and recalls of racial microaggressions specifically associated with Black individuals. Our aim was to provide an empirical understanding of the underlying themes, contexts, and linguistic patterns that distinguish acts from recollections. By shedding light on these distinctions, we contribute insights into the intricate nature of racial microaggressions and their manifestation in language. Our research uncovered significant distinctions between the acts and recalls of racial microaggressions, evident across a range of identified themes including Questions, Ability, Criminality, and more for acts, and Geographical Location, Everyday Life, Personality, and others for recalls. Furthermore, we delved deeper into their linguistic signatures through Critical Discourse Analysis (CDA), revealing more nuanced differences. In acts, we observed the utilization of absolute terminology, modifying adverbs, and statistical references, while recalls exhibited the prevalent use of past tense verbs and phrases such as “White people” and “Only Black.” Additionally, we identified commonalities in linguistic signatures between acts and recalls, such as the use of first-person language and the presence of Us vs. Them language. These shared linguistic patterns shed light on the underlying dynamics and interplay between the acts and recalls of racial microaggressions. Finally, this research serves as a foundation for further exploration and a catalyst for more informed discussions on addressing and combating racial microaggressions in both offline and online spaces.

6.1 Limitations

While our research is the first to systematically investigate acts and recalls of racial microaggressions comprehensively, our work is not without limitations. First, our analysis is limited to the context of racism in the U.S. Hence, implications around our findings may not be generalized to foreign contexts of racism against Blacks in other countries. Second, since Reddit is a global site, we have limited understanding of whether all the posters are from the U.S or not, which skews the earlier assumption of our analysis being limited to the context of racism in the U.S. Third, given the limited Black population in our area, we were only able to recruit Black college students for our workshop discussions. We plan to extend this study to a broader group in the future. Further, obtaining “ground-truth” labels for discursive data such as ours, are often fraught with subjective interpretations of race and social values linked with race-related matters, which are subject to the annotator’s own perspectives on racism, personal experiences, identity, and social background. Hence, while we endeavored towards obtaining “ground-truth” labels by discerning insights from discussions with Black participants through our workshop, we acknowledge that this process too, can be subject to biases.

Bibliography

[1] Social network sites as networked publics: Affordances, dynamics, and implications. In Zizi Papacharissi, editor, *A Networked Self*, pages 47–66. Routledge, 0 edition, 2010-09-10. ISBN 978-0-203-87652-7. doi: 10.4324/9780203876527-8.

[2] Kupiri Ackerman-Barger, Dowin Boatright, Rosana Gonzalez-Colaso, Regina Orozco, and Darin Latimore. Seeking inclusion excellence. *Academic Medicine*, 95(5):758–763, May 2020. doi: 10.1097/acm.0000000000003077. URL https://doi.org/10.1097/acm.0000000000003077.

[3] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Understanding the effect of deplatforming on social networks. In *13th ACM Web Science Conference 2021*, WebSci '21, pages 187–195. Association for Computing Machinery, 2021-06-21. ISBN 978-1-4503-8330-1. doi: 10.1145/3447535.3462637.

[4] Bobby Allyn. It's 'our fault': Nextdoor CEO takes blame for deleting of black lives matter posts, 1 Jul 2020. URL https://tinyurl.com/3353yeb4.

[5] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI Conference on Human Factors*

in Computing Systems, pages 3906–3918. ACM, 2016-05-07. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858096.

[6] Shervin Assari and Maryam Moghani Lankarani. Depressive symptoms and self-esteem in white and black older adults in the united states. *Brain sciences*, 8(6):105, 2018.

[7] James Banks. Regulating hate speech online. *International Review of Law, Computers & Technology*, 24(3):233–239, 2010.

[8] Daniel Bar-Tal. Self-censorship: The conceptual framework. *Self-censorship in contexts of conflict: Theory and research*, pages 1–18, 2017.

[9] Daniel Bar-Tal. Self-censorship as a socio-political-psychological phenomenon: Conception and research. *Political Psychology*, 38:37–65, 2017.

[10] Mario L. Barnes. *Law Society Review*, 49(1):279–282, 2015. ISSN 00239216, 15405893. URL http://www.jstor.org/stable/43670469.

[11] Roy F Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D Vohs. Bad is stronger than good. *Review of general psychology*, 5(4):323–370, 2001.

[12] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*, 2016.

[13] Su Lin Blodgett, Q. Vera Liao, Alexandra Olteanu, Rada Mihalcea, Michael Muller, Morgan Klaus Scheuerman, Chenhao Tan, and Qian Yang. Responsible Language Technologies: Foreseeing and Mitigating Harms. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391566. doi: 10.1145/3491101.3516502. URL https://doi.org/10.1145/3491101.3516502.

[14] Arthur W Blume, Laura V Lovato, Bryan N Thyken, and Natasha Denny. The relationship of microaggressions with alcohol use and anxiety among ethnic minority college students in a historically white institution. *Cultural diversity and ethnic minority psychology*, 18(1):45, 2012.

[15] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[16] Courtney M. Bonam, Hilary B. Bergsieker, and Jennifer L. Eberhardt. Polluting black space. 145:1561–1582, 2016. ISSN 1939-2222. doi: 10.1037/xge0000226. Place: US Publisher: American Psychological Association.

[17] Valeria Borsotti and Pernille Bjorn. Humor and stereotypes in computing: An equity-focused approach to institutional accountability. *Computer Supported Cooperative Work (CSCW)*, 31, 07 2022. doi: 10.1007/s10606-022-09440-9.

[18] Danah Boyd. Social network sites as networked publics: Affordances, dynamics, and implications. In *A networked self*, pages 47–66. Routledge, 2010.

[19] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674. Association for Computational Linguistics, 2019-11. doi: 10.18653/v1/D19-1176.

[20] Jerome S Bruner. On perceptual readiness. *Psychological review*, 64(2):123, 1957.

[21] Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.760.

[22] Nicoletta Cavazza and Margherita Guidetti. Swearing in political discourse: Why vulgarity works. *Journal of Language and Social Psychology*, 33(5):537–547, 2014.

[23] Stevie Chancellor, Zhiyuan (Jerry) Lin, and Munmun De Choudhury. "This Post Will Just Get Taken Down": Characterizing Removed Pro-Eating Disorder Social Media Content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 1157–1162, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858248. URL https://doi.org/10.1145/2858036.2858248.

[24] Stevie Chancellor, Yannis Kalantidis, Jessica A. Pater, Munmun De Choudhury, and David A. Shamma. Multimodal Classification of Moderated Online Pro-Eating Disorder Content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 3213–3226, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. doi: 10.1145/3025453.3025985. URL https://doi.org/10.1145/3025453.3025985.

[25] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25, 2018.

[26] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–30, 2019.

[27] Lijuan Chen, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. Toward the Understanding of Deep Text Matching Models for Information Retrieval. *arXiv preprint arXiv:2108.07081*, 2021.

[28] Elaine W Chun and Adrienne Lo. Language and racialization. In *The Routledge handbook of linguistic anthropology*, pages 220–233. Routledge, 2015.

[29] Chrissy Cook, Jie Cai, and Donghee Yvette Wohn. Awe Versus Aww: The Effectiveness of Two Kinds of Positive Emotional Stimulation on Stress Reduction for Online Content Moderators. *Proceedings of the ACM on Human-Computer Interaction*, 6:1 – 19, 2022.

[30] Juliet M. Corbin and Anselm Strauss. Basics of qualitative research (3rd ed.): Techniques and procedures for developing grounded theory. 2008.

[31] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 699–708, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312295. doi: 10.1145/2187836.2187931. URL https://doi.org/10.1145/2187836.2187931.

[32] Anubrata Das, Brandon Dang, and Matthew Lease. Fast, accurate, and healthier: Interactive blurring helps moderators reduce exposure to harmful content. In *AAAI Conference on Human Computation & Crowdsourcing*, 2020.

[33] Richard Delgado and Jean Stefancic. Why do we tell the same stories?: Law reform,

critical librarianship, and the triple helix dilemma. *Stanford Law Review*, pages 207–225, 1989.

[34] Richard Delgado and Jean Stefancic. Images of the outsider in American law and culture: Can free expression remedy systemic social ills. *Cornell L. Rev.*, 77:1258, 1991.

[35] Richard Delgado and Jean Stefancic. Critical race theory: An annotated bibliography. *Virginia Law Review*, 79(2):461–516, 1993. ISSN 00426601. URL http://www.jstor.org/stable/1073418.

[36] Nicola Döring and M Rohangis Mohseni. Male dominance and sexism on youtube: results of three content analyses. *Feminist Media Studies*, 19(4):512–524, 2019.

[37] Bryan Dosono. AAPI Identity Work on Reddit: Toward Social Support and Collective Action. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, GROUP '18, page 373–378, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355629. doi: 10.1145/3148330.3152697. URL https://doi.org/10.1145/3148330.3152697.

[38] Bryan Dosono and Bryan Semaan. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300372. URL https://doi.org/10.1145/3290605.3300372.

[39] Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. Towards interpreting and mitigating shortcut learning behavior of NLU models. *arXiv preprint arXiv:2103.06922*, 2021.

[40] Maeve Duggan. 1 in 4 Black Americans have faced online harassment because of their race or ethnicity, 25 Jul 2017. URL https://tinyurl.com/5n8n845e.

[41] Jr. Dulaney, Earl F. Changes in Language Behavior as a Function of Veracity. *Human Communication Research*, 9(1):75–82, 03 2006. ISSN 0360-3989. doi: 10.1111/j.1468-2958.1982.tb00684.x. URL https://doi.org/10.1111/j.1468-2958.1982.tb00684.x.

[42] Anca Dumitrache, Oana Inel, Benjamin Timmermans, Carlos Ortiz, Robert-Jan Sips, Lora Aroyo, and Chris Welty. Empirical methodology for crowdsourcing ground truth. *Semantic Web*, 12(3):403–421, 2021.

[43] Elizabeth Dwoskin, Nitasha Tiku, and Craig Timberg. Facebook's race-blind practices around hate speech came at the expense of black users, new documents show, Nov 2021. URL https://www.washingtonpost.com/technology/2021/11/21/facebook-algorithm-biased-race/.

[44] Sara E Shaw and Julia Bailey. Discourse analysis: what is it and why is it relevant to family practice? *Family Practice*, 26(5):413–419, 06 2009. ISSN 0263-2136. doi: 10.1093/fampra/cmp038. URL https://doi.org/10.1093/fampra/cmp038.

[45] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. Expanding explainability: Towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445188. URL https://doi.org/10.1145/3411764.3445188.

[46] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods*

in Natural Language Processing, pages 345–363. Association for Computational Linguistics, 2021-11. doi: 10.18653/v1/2021.emnlp-main.29.

[47] Rob Eschmann. Digital resistance: How online communication facilitates responses to racial microaggressions. *Sociology of Race and Ethnicity*, 7(2):264–277, 2021. doi: 10.1177/2332649220933307. URL https://doi.org/10.1177/2332649220933307.

[48] Rob Eschmann, Jacob Groshek, Rachel Chanderdatt, Khea Chang, and Maysa Whyte. Making a microaggression: Using big data and qualitative analysis to map the reproduction and disruption of microaggressions through social media. *Social Media + Society*, 6(4):2056305120975716, 2020. doi: 10.1177/2056305120975716. URL https://doi.org/10.1177/2056305120975716.

[49] Andre Espaillat, Danielle Panna, Dianne L. Goede, Matthew J. Gurka, Maureen A Novak, and Zareen Zaidi. An exploratory study on microaggressions in medical school: What are they and why should we care? *Perspectives on Medical Education*, 8:143 – 151, 2019.

[50] Philomena Essed. *Understanding everyday racism: An interdisciplinary theory.* Understanding everyday racism: An interdisciplinary theory. Sage Publications, Inc, 1991. ISBN 978-0-8039-4255-4 978-0-8039-4256-1. Pages: x, 322.

[51] Allyson Ettinger. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 2020.

[52] Jessica L Feuston, Alex S Taylor, and Anne Marie Piper. Conformity of eating disorders through content moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–28, 2020.

[53] Meira Gebel. Black creators say Tiktok still secretly hides their content, 21 Jul 2020. URL https://www.digitaltrends.com/social-media/black-creators-claim-tiktok-still-secretly-blocking-content/.

[54] James Paul Gee. *An Introduction to Discourse Analysis: Theory and Method*. routledge, 2014.

[55] Arpita Ghosh, Satyen Kale, and Preston McAfee. Who moderates the moderators? crowdsourcing abuse detection in user-generated content. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, EC '11, page 167–176, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302616. doi: 10.1145/1993574.1993599. URL https://doi.org/10.1145/1993574.1993599.

[56] Graham R Gibbs. *Analyzing qualitative data*, volume 6. Sage, 2018.

[57] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[58] Tarleton Gillespie. Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, 8(3):20563051221117552, 2022.

[59] Phillip Atiba Goff, Jennifer L Eberhardt, Melissa J Williams, and Matthew Christian Jackson. Not yet human: implicit knowledge, historical dehumanization, and contemporary consequences. *Journal of personality and social psychology*, 94(2):292, 2008.

[60] Joseph P Goodwin. *Stories in the time of cholera: Racial profiling during a medical nightmare*. JSTOR, 2002.

[61] Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance.

Big Data & Society, 7(1):2053951719897945, 2020. doi: 10.1177/2053951719897945. URL https://doi.org/10.1177/2053951719897945.

[62] Jonathan Grudin. Why CSCW applications fail: problems in the design and evaluationof organizational interfaces. In *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*, pages 85–93, 1988.

[63] Jessica Guynn. Facebook while black: Users call it getting 'Zucked' say talking about racism is censored as hate speech. *Usa today*, 24, July 2019.

[64] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021. doi: 10.1145/3479610. URL https://doi.org/10.1145/3479610.

[65] Joanne M Hall and Becky Fields. "It's killing us!" Narratives of Black adults about microaggression experiences and related health stress. *Global qualitative nursing research*, 2:2333393615591569, 2015.

[66] Chayla Haynes, Saran Stewart, and Evette Allen. Three paths, one struggle: Black women and girls battling invisibility in U.S classrooms. *Journal of Negro Education*, 85(3):380–391, 2016.

[67] Danula Hettiachchi and Jorge Goncalves. Towards effective crowd-powered online content moderation. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, OZCHI'19, page 342–346, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450376969. doi: 10.1145/3369457.3369491. URL https://doi.org/10.1145/3369457.3369491.

[68] Sharon Heung, Mahika Phutane, Shiri Azenkot, Megh Marathe, and Aditya Vashistha. Nothing micro about it: Examining ableist microaggressions on social media. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392587. doi: 10.1145/3517428.3544801. URL https://doi.org/10.1145/3517428.3544801.

[69] Jon Hurwitz and Mark Peffley. Public perceptions of race and crime: The role of racial stereotypes. *American Journal of Political Science*, 41:375, 04 1997. doi: 10.2307/2111769.

[70] Jevan A. Hutson, Jessie G. Taft, Solon Barocas, and Karen Levy. Debiasing desire: Addressing bias discrimination on intimate platforms. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), Nov 2018. doi: 10.1145/3274342. URL https://doi.org/10.1145/3274342.

[71] Miyako Inoue. Word for word: Verbatim as political technologies. *Annual Review of Anthropology*, 47:217–232, 2018.

[72] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2):1–33, 2018.

[73] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. "Did You Suspect the Post Would Be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019. doi: 10.1145/3359294. URL https://doi.org/10.1145/3359294.

[74] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does transparency in moderation

really matter? user behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27, 2019.

[75] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–30, 2021.

[76] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. Through the looking glass: Study of transparency in reddit's moderation practices. *Proceedings of the ACM on Human-Computer Interaction*, 4(GROUP):1–35, 2020.

[77] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376219. URL https://doi.org/10.1145/3313831.3376219.

[78] Micere Keels, Myles Durkee, and Elan Hope. The psychological and academic costs of school-based racial and ethnic microaggressions. *American Educational Research Journal*, 54(6):1316–1344, 2017.

[79] Kennedy Kelis. Deconstructing the "13/50" argument, 10 Oct 2021. URL https://detester.org/publications/162kennedy.

[80] Ben Kew. Poll: Two-thirds of Conservatives don't trust Facebook, believe social media censors conservatives. *Breitbart*, 29 Aug 2018. URL https://tinyurl.com/pzavznpb.

[81] Gohar Feroz Khan, Bobby Swar, and Sang Kon Lee. Social media risks and benefits:

A public sector perspective. 32(5):606–627, 2014-10-01. ISSN 0894-4393. doi: 10.1177/0894439314524701. Publisher: SAGE Publications Inc.

[82] Yubo Kou and Xinning Gui. Mediating Community-AI Interaction through Situated Explanation: The Case of AI-Led Moderation. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), oct 2020. doi: 10.1145/3415173. URL https://doi.org/10.1145/3415173.

[83] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3501999. URL https://doi.org/10.1145/3491102.3501999.

[84] J Richard Landis and Gary G Koch. An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. *Biometrics*, pages 363–374, 1977.

[85] Alyssa Lees, Daniel Borkan, Ian Kivlichan, Jorge Nario, and Tesh Goyal. Capturing covertly toxic speech via crowdsourcing. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 14–20. Association for Computational Linguistics, 2021-04.

[86] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. *Philosophy Technology*, 31, 12 2018. doi: 10.1007/s13347-017-0279-x.

[87] Weiwen Leung, Zheng Zhang, Daviti Jibuti, Jinhao Zhao, Maximilian Klein, Casey Pierce, Lionel Robert, and Haiyi Zhu. Race, gender and beauty: The effect of information provision on online hiring biases. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–11, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376874. URL https://doi.org/10.1145/3313831.3376874.

[88] Daniel Link, Bernd Hellingrath, and Jie Ling. A human-in-the-loop approach for semi-automated content moderation. In *International Conference on Information Systems for Crisis Response and Management*, 2016.

[89] Yang Liu, Christopher Whitfield, Tianyang Zhang, Amanda Hauser, Taeyonn Reynolds, and Mohd Anwar. Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning. *Health Information Science and Systems*, 9(1), June 2021. doi: 10.1007/s13755-021-00158-4. URL https://doi.org/10.1007/s13755-021-00158-4.

[90] Andrea Marshall, Angela D. Pack, Sarah A. Owusu, Rainbo Hultman, David Drake, Florentine U. N. Rutaganira, Maria Namwanje, Chantell S. Evans, Edgar Garza-Lopez, Samantha C. Lewis, Cristina Termini, Salma AshShareef, Innes Hicsasmaz, Brittany L. Taylor, Melanie R. McReynolds, Haysetta D. Shuler, and Antentor O. Hinton. Responding and navigating racialized microaggressions in stem. *Pathogens and Disease*, 79, 2021.

[91] Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. Hate Begets Hate: A Temporal Study of Hate Speech. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), oct 2020. doi: 10.1145/3415163. URL https://doi.org/10.1145/3415163.

[92] Ryan Moore. *The New Jim Crow*. Macat Library, July 2017. doi: 10.4324/9781912282586. URL https://doi.org/10.4324/9781912282586.

[93] Samantha L Moore-Berg, Boaz Hameiri, and Emile G Bruneau. Empathy, dehumanization, and misperceptions: A media intervention humanizes migrants and increases empathy for their plight but only if misinformation about migrants is also corrected. *Social Psychological and Personality Science*, 13(2):645–655, 2022.

[94] Michael Muller and Thomas Erickson. In the Data Kitchen: A Review (a Design Fiction on Data Science). In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, page 1–10, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356213. doi: 10.1145/3170427.3188407. URL https://doi.org/10.1145/3170427.3188407.

[95] Michael Muller, Christine T. Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, and Casey Dugan. Designing ground truth and the social life of labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445402. URL https://doi.org/10.1145/3411764.3445402.

[96] Kevin L Nadal, Katie E Griffin, Yinglee Wong, Sahran Hamit, and Morgan Rasmus. The impact of racial microaggressions on mental health: Counseling implications for clients of color. *Journal of Counseling & Development*, 92(1):57–66, 2014.

[97] Kate G Niederhoffer and James W Pennebaker. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360, 2002.

[98] Poppy Noor. The celebrities who are doing anti-racism right, 15 Jun 2020. URL https://www.theguardian.com/world/2020/jun/15/selena-gomez-anti-racism-celebrities-getting-it-right.

[99] Fayika Farhat Nova, MD. Rashidujjaman Rifat, Pratyasha Saha, Syed Ishtiaque Ahmed, and Shion Guha. Online Sexual Harassment over Anonymous Social Media in Bangladesh. In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development*, ICTD '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361224. doi: 10.1145/3287098.3287107. URL https://doi.org/10.1145/3287098.3287107.

[100] Fayika Farhat Nova, Michael Ann DeVito, Pratyasha Saha, Kazi Shohanur Rashid, Shashwata Roy Turzo, Sadia Afrin, and Shion Guha. "Facebook Promotes More Harassment" Social Media Ecosystem, Skill and Marginalized Hijra Identity in Bangladesh. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–35, 2021.

[101] Ihudiya Finda Ogbonnaya-Ogburu, Angela D.R. Smith, Alexandra To, and Kentaro Toyama. Critical race theory for HCI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–16. ACM, 2020-04-21. ISBN 978-1-4503-6708-0. doi: 10.1145/3313831.3376392.

[102] Jason A Okonofua, Gregory M Walton, and Jennifer L Eberhardt. A vicious cycle: A social–psychological account of extreme racial disparities in school discipline. *Perspectives on Psychological Science*, 11(3):381–398, 2016.

[103] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User*

Interface Software and Technology, UIST '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393201. doi: 10.1145/3526113.3545616. URL https://doi.org/10.1145/3526113.3545616.

[104] Joon Sung Park, Joseph Seering, and Michael S. Bernstein. Measuring the prevalence of anti-social behavior in online communities. *Proceedings of the ACM on Human-Computer Interaction*, 6:1 – 29, 2022.

[105] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

[106] Jennifer L. Eberhardt PhD. *Biased: Uncovering the Hidden Prejudice That Shapes What We See, Think, and Do*. Penguin, 2020-03-03. ISBN 978-0-7352-2495-7. Google-Books-ID: hpDODwAAQBAJ.

[107] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5740–5745. Association for Computational Linguistics, 2019-11. doi: 10.18653/v1/D19-1578.

[108] Emilee Rader, Kelley Cotter, and Janghee Cho. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173677. URL https://doi.org/10.1145/3173574.3173677.

[109] Aneeta Rattan and Jennifer L Eberhardt. The role of social meaning in inattentional blindness: When the gorillas in our midst do not go unseen. *Journal of Experimental Social Psychology*, 46(6):1085–1088, 2010.

[110] Elizabeth Reid, Regan L. Mandryk, Nicole A. Beres, Madison Klarkowski, and Julian Frommel. Feeling good and in control: In-game tools to support targets of toxicity. *Proc. ACM Hum.-Comput. Interact.*, 6(CHI PLAY), oct 2022. doi: 10.1145/3549498. URL https://doi.org/10.1145/3549498.

[111] Angela Reyes. *Language, identity, and stereotype among Southeast Asian American youth: The other Asian*. Routledge, 2017.

[112] Eugenia Ha Rim Rho and Melissa Mazmanian. Hashtag burnout? a control experiment investigating how political hashtags shape reactions to news content. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–25, 2019.

[113] Eugenia Ha Rim Rho and Melissa Mazmanian. Political hashtags & the lost art of democratic discourse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM, 2020-04-21. ISBN 978-1-4503-6708-0. doi: 10.1145/3313831.3376542.

[114] Eugenia Ha Rim Rho, Oliver L. Haimson, Nazanin Andalibi, Melissa Mazmanian, and Gillian R. Hayes. Class confessions: Restorative properties in online experiences of socioeconomic stigma. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3377–3389. Association for Computing Machinery, 2017-05-02. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3025921.

[115] Eugenia Ha Rim Rho, Gloria Mark, and Melissa Mazmanian. Fostering civil discourse online: Linguistic behavior in comments of# metoo articles across political perspectives. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–28, 2018.

[116] Travis Riddle and Stacey Sinclair. Racial disparities in school-based disciplinary actions are associated with county-level rates of racial bias. *Proceedings of the National Academy of Sciences*, 116(17):8255–8260, 2019.

[117] Sarah Roberts. Digital detritus: 'error' and the logic of opacity in social media content moderation. *First Monday*, 23, 03 2018. doi: 10.5210/fm.v23i3.8283.

[118] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678. Association for Computational Linguistics, 2019-07. doi: 10.18653/v1/P19-1163.

[119] Morgan Klaus Scheuerman, Stacy M. Branham, and Foad Hamidi. Safe spaces and safe places: Unpacking technology-mediated experiences of safety and harm with transgender people. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), nov 2018. doi: 10.1145/3274424. URL https://doi.org/10.1145/3274424.

[120] Christoph Schneider, Markus Weinmann, and Jan Vom Brocke. Digital nudging: guiding online user choices through interface design. *Communications of the ACM*, 61(7): 67–73, 2018.

[121] Daniel J Simons and Christopher F Chabris. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *perception*, 28(9):1059–1074, 1999.

[122] C. Estelle Smith, William Lane, Hannah Miller Hillberg, Daniel Kluver, Loren Terveen, and Svetlana Yarosh. Effective Strategies for Crowd-Powered Cognitive Reappraisal Systems: A Field Deployment of the Flip*Doubt Web Application for Mental Health. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021. doi: 10.1145/3479561. URL https://doi.org/10.1145/3479561.

[123] Michael Spencer. Microaggressions and social work practice, education, and research. *Journal of Ethnic Cultural Diversity in Social Work*, 26:1–5, 02 2017. doi: 10.1080/15313204.2016.1268989.

[124] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445092. URL https://doi.org/10.1145/3411764.3445092.

[125] Stefan Stieglitz and Linh Dang-Xuan. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems*, 29(4):217–248, 2013.

[126] Derald Wing Sue and Lisa Beth Spanierman. *Microaggressions in everyday life, 2nd ed*. Microaggressions in everyday life, 2nd ed. John Wiley & Sons, Inc., 2020. ISBN 978-1-119-51379-7 978-1-119-51380-3 978-1-119-51381-0. Pages: xx, 349.

[127] Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha Holder, Kevin L Nadal, and Marta Esquilin. Racial microaggressions in everyday life: implications for clinical practice. *American psychologist*, 62(4):271, 2007.

[128] Derald Wing Sue, Sarah Alsaidi, Michael N Awad, Elizabeth Glaeser, Cassandra Z Calle, and Narolyn Mendez. Disarming racial microaggressions: Microintervention strategies for targets, White allies, and bystanders. *American Psychologist*, 74(1):128, 2019.

[129] S Shyam Sundar, Haiyan Jia, T Franklin Waddell, and Yan Huang. Toward a theory of interactive media effects (TIME) four models for explaining how interface features

affect user psychology. *The handbook of the psychology of communication technology*, pages 47–86, 2015.

[130] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[131] Nicolas Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. What do we mean when we talk about transparency? Towards meaningful transparency in commercial content moderation. *International Journal of Communication*, 13:1526–1543, 2019. URL https://eprints.qut.edu.au/126386/.

[132] Michael Swart, Ylana Lopez, Arunesh Mathur, and Marshini Chetty. Is This An Ad?: Automatically Disclosing Online Endorsements On YouTube With AdIntuition. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376178. URL https://doi.org/10.1145/3313831.3376178.

[133] Yla R. Tausczik and James W. Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010. doi: 10.1177/0261927X09351676. URL https://doi.org/10.1177/0261927X09351676.

[134] Hibby Thach, Samuel Mayworm, Daniel Delmonaco, and Oliver Haimson. (In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. *New Media & Society*, 0(0):14614448221109804, 0. doi: 10.1177/14614448221109804. URL https://doi.org/10.1177/14614448221109804.

[135] Roelien C. Timmer, David Liebowitz, Surya Nepal, and Salil S. Kanhere. Can pre-trained Transformers be used in detecting complex sensitive sentences? - a Monsanto case study. In *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 90–97, 2021-12. doi: 10.1109/TPSISA52974.2021.00010.

[136] Alexandra To, Wenxia Sweeney, Jessica Hammer, and Geoff Kaufman. “They Just Don’t Get It”: Towards Social Technologies for Coping with Interpersonal Racism. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–29, 2020.

[137] Alexandra To, Hillary Carey, Geoff Kaufman, and Jessica Hammer. Reducing uncertainty and offering comfort: Designing technology for coping with interpersonal racism. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445590. URL https://doi.org/10.1145/3411764.3445590.

[138] Stephen M Utych. Negative affective language in politics. *American Politics Research*, 46(1):77–102, 2018.

[139] Sonja Utz and Johannes Breuer. Informational benefits from social media use for professional purposes: Results from a longitudinal study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 10(4), 2016.

[140] Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI ’18, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery.

ISBN 9781450356206. doi: 10.1145/3173574.3174014. URL https://doi.org/10.1145/3173574.3174014.

[141] James Vincent. Nextdoor CEO says it's 'our fault' moderators deleted black lives matter posts, 2 Jul 2020. URL https://www.theverge.com/2020/7/2/21311046/nextdoor-ceo-admits-fault-moderators-racial-bias-black-lives-matter.

[142] Kelly Welch. Black criminal stereotypes and racial profiling. *Journal of Contemporary Criminal Justice*, 23(3):276–288, 2007. doi: 10.1177/1043986207306870. URL https://doi.org/10.1177/1043986207306870.

[143] Sarah Myers West. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383, 2018. doi: 10.1177/1461444818773059. URL https://doi.org/10.1177/1461444818773059.

[144] Monnica T Williams, Matthew D Skinta, and Renée Martin-Willett. After pierce and sue: A revised racial microaggressions taxonomy. *Perspectives on Psychological Science*, 16(5):991–1007, 2021.

[145] Donghee Yvette Wohn. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300390. URL https://doi.org/10.1145/3290605.3300390.

[146] Austin Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Duen Chau, and Diyi Yang. RECAST: Enabling User Recourse and In-

terpretability of Toxicity Detection Models with Interactive Visualization. *Proceedings of the ACM on Human-Computer Interaction*, 5:1–26, 04 2021. doi: 10.1145/3449280.

[147] Austin P Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Diyi Yang, and Duen Horng Chau. Recast: Interactive auditing of automatic toxicity detection models. In *The Eighth International Workshop of Chinese CHI*, Chinese CHI 2020, page 80–82, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450388153. doi: 10.1145/3403676.3403691. URL https://doi.org/10.1145/3403676.3403691.

[148] Qunfang Wu, Louisa Kayah Williams, Ellen Simpson, and Bryan Semaan. Conversations About Crime: Re-Enforcing and Fighting Against Platformed Racism on Reddit. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), apr 2022. doi: 10.1145/3512901. URL https://doi.org/10.1145/3512901.

[149] Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. Exploring Antecedents and Consequences of Toxicity in Online Discussions: A Case Study on Reddit. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), oct 2020. doi: 10.1145/3415179. URL https://doi.org/10.1145/3415179.

[150] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376301. URL https://doi.org/10.1145/3313831.3376301.

[151] Savvas Zannettou, Mai ElSherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca

Stringhini. Measuring and characterizing hate speech on news websites. In *12th ACM Conference on Web Science*, pages 125–134, 2020.

Appendices

Appendix A

Annotation Guidelines for Workshop Participants

A.1 Appendix

We analyze 13 of 16 themes of racial microaggressions from the revised Sue et al. (2009)'s taxonomy [144]. Three of the 16 themes (tokenism, environmental exclusion, and environmental attacks) were excluded as they were not applicable to online contexts: the three themes pertain to situations in which people are present in a physical environment. For each of the 13 themes, we provide the definition from [144] and examples of racial microaggressions from our RAMA corpus in Table A.1. We discussed these examples and definitions with workshop participants as an annotation guideline for labeling instances of acts of online racial microaggressions.

Theme	Definition	Examples
THEME-1: Alien in own land / Not a True Citizen	When a question, statement, or behavior indicates that a person of color is not a real citizen or a meaningful part of society because they are not White; Questioning the legitimacy of their identity.	•"If Adam and Eve are the first people in the Earth and they are white, why are there Black people?" •"God gave black people rights in all corners of the globe and then he made the Earth round."
THEME-2: Racial Categorization and Sameness	When a person is compelled to disclose their racial group to enable others to attach pathological racial stereotypes to the person; includes the assumption that all people from a particular group are alike	•"All Black people look alike."
THEME-3: Assumptions about intelligence, competence, or status	When behavior or statements are based on an assumption about a person's intelligence, competence, education, income, or social status derived from racial stereotypes.	•"How did Black people get so good at science? and why are they so athletic?"
THEME-4: Connecting via stereotypes	When a person tries to communicate or connect with a person through the use of stereotyped speech or behavior to be accepted or understood; can include racist jokes and epitaphs as terms of endearment.	•"Why do Black people love fried chicken and watermelon?" •"Why don't Black people tip?"
THEME-5: False color blindness/ invalidating racial or ethnic identity	Expressing that an individual's racial or ethnic identity should not be acknowledged, which can be invalidating for people who are proud of their identity or who have suffered because of it.	•"I don't care if they are black, gay, green objects." •"I'm white and I don't care, we're all the same human race."
THEME-6: Myth of meritocracy/ race is irrelevant for success	When someone makes statements about success being rooted in personal efforts and denial of the existence of racism/White privilege; Statements which assert that race does not play a role in succeeding in career advancement or education.	•"Role should go to the best performer regardless of race." •"Rich black people don't face any form of systematic racism."
THEME-7: Reverse-racism hostility	Expressions of jealousy or hostility surrounding the notion that POC get unfair advantages and benefits because of their race.	•"I was fully qualified for the job, but they gave it to a Black girl." •"Oh wait you're Black; they practically guarantee you'd get into that college."
THEME-8: Criminality or dangerousness	Demonstrating belief in stereotypes that POC are dangerous, untrustworthy, and likely to commit crimes or cause bodily harm; A person of color is presumed to be dangerous, criminal, or deviant on the basis of their race.	•"I held back because he was Black" [user is speaking in the context of avoiding conflict with a Black person out of fear of physical retaliation]. •"Black men are dangerous."
THEME-9: Avoidance and distancing	When POC are avoided or measures are taken to prevent physical contact or close proximity.	•"When I see a Black person approaching me, I cross the street."
THEME-10: Denial of individual racism	When a person tries to make a case that they are not biased, often by talking about antiracist things they have done to deflect perceived scrutiny of their own biased behaviors; A statement made when Whites renounce their racial biases.	•"I'm not a racist. I have several Black friends." •"I'm not racist but Black people make me uncomfortable."
THEME-11: Pathologizing minority culture or appearance	When people criticize others on the basis of perceived or real cultural differences in appearance, traditions, behaviors, or preferences; The notion that the values and communication styles of the dominant culture are ideal.	• "Black kids shouldn't dress that way." • You're pretty for a Black girl.
THEME-12: Exoticization and eroticization	When a person of color is treated according to sexualized stereotypes or attention to differences that are characterized as exotic in some way.	•"Black women are exotic. I have a fetish for Black women, am I racist?"
THEME-13: Second-class citizen/ ignored and invisible	When POC are treated with less respect, consideration, or care than is normally expected or customary; may include being ignored or being unseen/ invisible; Occurs when a White person is given preferential treatment as a consumer over a person of color.	•"Oh, sorry we kept you waiting so long. From your surname on the form, we thought you were Black!" •"Black people and LGBT are untouchable."

Table A.1: Themes and examples of acts of racial microaggressions from our RAMA corpus.