Understanding Self-Narration of Personally Experienced Racism on Reddit

Diyi Yang,[1] **Scott Counts**[2]
[1] Language Technologies Institute, Carnegie Mellon University diyiy@cs.cmu.edu
[2] Microsoft Research counts@microsoft.com

Abstract

We identify and classify users self-narration of racial discrimination and corresponding community support in social media. We developed natural language models first to distinguish self-narration of racial discrimination in Reddit threads, and then to identify which types of support are provided and valued in subsequent replies. Our classifiers can detect the self-narration of personally experienced racism in online textual accounts with 83% accuracy, and can recognize four types of supportive actions in replies with up to 88% accuracy. Descriptively, our models identify types of racism experienced and the racist concepts (e.g., sexism, appearance or accent related) most experienced by people of different races. Finally we show that commiseration is the most valued form of social support.

Introduction

Racism has been identified as a social determinant of health (Mays, Cochran, and Barnes 2007; Paradies et al. 2013), and the experience of racial discrimination is strongly associated with poor health outcomes, both mental and physical, across diverse minority groups (Clark et al. 1999). As many social interactions move online, detecting and understanding racism in social media becomes increasingly important, and in fact studies show that racism is prevalent online (Chaudhry 2015; Melican and Dixon 2008). One recent report (Bartlett et al. 2014) showed that on average, roughly 10,000 people per day issue racist and ethnic slurs on Twitter. Minimizing racism online may be critical in the longer term given the rise of artificial intelligence-driven natural language interaction systems that learn from large scale conversation and other text corpora drawn from a wide cross-section of society. Fortunately, these same social media platforms contain data that we can leverage proactively to build systems for automatic detection of racism and potentially to trigger support systems.

In this work, we characterize a specific instance of racism - self-narration of personally experienced racism that people express on Reddit. **Self-narrations** of experienced racism are victims' direct disclosure of their negative experiences. For example, a user posted this to Reddit: *"I'm a nurse. I walked into this lady's hospital room to start my assessment. She looks at me, hits the call button and asks for the charge nurse. Mind you I have only gone as far as introducing myself. She looks to the charge nurse and says, "I want a different nurse. One who is smart and knows what they are doing. She's black, she can't possibly know what she is doing" "*. Thousands of such self-narrations of racial discrimination can provide the scale of data needed to capture the wide range of people's experiences with discrimination, along with their subsequent thoughts and feelings.

By self-disclosing their experiences of being discriminated, users might expect to receive understanding or help from their fellow community members. To that end, Reddit threads on discrimination also contain supportive responses from other users that can help us start to leverage technological systems to better understand and help support those experiencing racism. These responses could be about sharing a similar experience "*The same thing happened to me when I was ...*" or understanding and feeling what another person is experiencing "*I totally understand what you're saying...*". Previous research demonstrates that such support can mitigate the deleterious impact of racial discrimination on health (Seawell, Cutrona, and Russell 2012). While the Reddit community itself can provide some of this support, not every thread receives supportive replies in a timely fashion, and of course other online environments may have little to no supportive community. A large scale characterization of experiences of racism and related social support can provide social and health scientists, practitioners, and caregivers a better sense of the scope and nuance of racism as well as approaches to supporting those suffering discrimination.

In this work, we developed a high accuracy text classifier to identify posts that contain users' self-narration of personally experienced racism (versus discussing of racism news). We then used it to describe the nuances of various forms of racism experienced, such as an extraction of the common linguistic topics in descriptions of racism experienced, and recognized different types of community support and understanding the effectiveness of different forms of support.

Related Work

Only a few studies have investigated racial discrimination in social media. For example, De Choudhury et al. 2016 looked at the activist movement "Black Lives Matter" around racial discrimination and police violence, finding that the affective, behavioral and linguistic measures derived from social media can predict future protest participation on the ground.

Related, Olteanu et al. 2015 studied the demographics of Twitter users who used the hashtag #BlackLivesMatter and found blacks to be more engaged with the hashtag. Instead of focusing on a specific social movement, our work tries to detect people's self-narration of their personal experiences of racial discrimination and then analyze how the community responds to them in a supportive way. Methodologically, using naturalistic data to study racial discrimination requires a critical first step of identifying instances of discrimination with sufficient accuracy. Traditional studies on measuring racial discrimination have used surveys to ask racial minorities about their experiences with discrimination in social settings, or have developed interview techniques aimed at gauging propensities toward discrimination among the general population (Schuman 1997). Identifying racism in social media can be difficult given the scale and unstructured nature of data. In our work, we built machine learning models to distinguish accounts of personally experienced racism at scale based on linguistic evidences in social media posts.

Ethical Considerations Racism is a sensitive area, and research on the topic requires careful attention to ethical issues. In terms of data, we used only public data (de-identified) from Reddit. In terms of exposing people to sensitive content, Mechanical Turk workers were shown sample sets of the Reddit posts and replies in order to provide labels on the post text. These workers were made aware of the content topic area in the task title and instructions, were free to skip any item, and could stop labeling at any time.

Dataset and Measures

Our analysis is conducted on Reddit threads and their replies. To acquire a diverse and representative dataset about candidates of self-narrations of racial discrimination, we collected messages from the subreddit *racism* as well as messages from the subreddit *AskReddit* that contained relevant keywords such as 'racism'. Because many top level threads in Reddit contain replies which themselves contain many accounts of racism experiences, we also treated direct replies to the original threads as "threads". Empty or deleted messages were removed from our corpus. All messages in our sample appeared before June 2016, and to control for length effects, contain between 200 and 1500 characters. This resulted in 3146 threads and 5212 direct replies in total. Note that some victims might describe experiences of being discriminated that happened to them this week, while others were describing things that they experienced a long time ago. We did not differentiate them; instead, we hope our detector can detect any instances of self-narrations immediately once users self-narrated their personally experienced racism online, and then inform possible helpers from the community to provide appropriate support.

Data Annotation We used Amazon's Mechanical Turk (MTurk) to construct a reliable, hand-coded dataset to measure whether a user actually self-narrated an experience of racial discrimination. Given a thread, we asked Turkers to objectively complete two tasks: (1) judge whether the author is describing his/her experience of being discriminated against by others and how much he/she is negatively affected by it, and (2) rate which race or ethnicity is being discriminated against. For (1), Turkers could choose from 'No: the author is not describing a personal experience of discrimination', 'Yes: not affected', 'Yes: slightly affected', 'Yes: moderately affected' and 'Yes: seriously affected'. For (2), we asked Turkers to choose from 'East Asian', 'South Asian', 'African-American or Black', 'White', 'Hispanic/Latino', 'Native American', and 'Native Hawaiian or Other Pacific Islander'. To increase annotation quality, we required Turkers to have a United States location and 95% approval rate for their previous work on MTurk. We randomly selected 800 threads from our dataset. Each thread was labeled by two Master workers[1]. Turkers received $0.08 for rating each thread. In terms of the reliability of annotations, across all annotations on whether a user is describing an experience of racial discrimination, we obtained an Intra-class Correlation (ICC) score of 0.736, and an ICC of 0.933 for which race is discriminated, ICC is 0.933 (Cicchetti 1994).

	Acc	**Recall**	**Prec**	**F1**
Racial Discrimination	0.830	0.874	0.913	0.892
Severity	0.457	0.369	0.457	0.403
Against Which Race	0.659	0.634	0.659	0.606

Table 1: Classification performance for predicting whether a user is describing a personal experience of racial discrimination (Racial Discrimination), how much the author is affected by it (Degree of Affect) and which race is discriminated against (Against Which Race).

Measures We consider a number of measures for identifying and characterizing self-narrations of racial discrimination. **Affective Expression**: Descriptions of negative experiences of racism often contain words carrying strong sentiment. To identify this word-emotion association, we computed several affect-relevant measures using LIWC: *positive emotion*, *negative emotion*, *anger, anxiety* and *sadness*. **Interpersonal Pronoun**: People might refer to experiences or interactions related to themselves about racial discrimination using personal pronouns, such as 'I', 'them', 'we'.We assess this interpersonal reflection by measuring the frequency of usage of *first-personal singular, first-personal plural, second person, third-person, third-person singular, third-person plural* and *articles*. **Spoken Language**: Informal language might reveal people's surprise or anger in recounting their experiences of racism. Thus we measure the frequency of *swear* words such as 'damn', 'fuck'. **Social Concerns**: Self-narration of racial discrimination might happen when interacting with others or others are involved in everyday social settings. To capture this social process, we assessed the occurrences of words in the categories of *social*, *family* ('parents', 'daughter'), *friends* ('friend', 'neighbor') and *humans* ('adult', 'kids'). **Bag of Words**: To capture the racism contained in messages, we also considered the predictive effect contributed by each unigram.

[1] https://www.mturk.com/mturk/help?helpPage=worker\ #what_is_master_worker

	Accuracy	Recall	Precision	F1
Commiseration	0.883	0.743	0.858	0.795
Empathy	0.885	0.365	0.680	0.474
Info Support	0.745	0.680	0.786	0.726
Encouragement	0.775	0.760	0.802	0.769
No Support	0.821	0.537	0.589	0.560

Table 2: Performance for classifying the provision of different types of social support. Chance accuracy is 50%. The majority baseline accuracies are 69.2%, 85.5%, 50.0%, 50.0% and 78.6% for commiseration, empathy, info support, encouragement, and no support respectively.

Characterizing Experienced Racism

In this part, we build language-based classifiers that can identify content about people's *self-narration of personal experiences with racism*, versus for example the discussion of news events, and then how impactful was the experience. We only kept threads that had unanimous labels across the two MTurk judges. This resulted in 660 threads, among which 508 were labeled as experiencing racial discrimination. All these text classification experiments incorporated the linguistic measures and bag of word features, and were evaluated with 10-fold cross validation as shown in Table 1. The by-chance baseline accuracies for three tasks are 50%, 25% and 16.7% respectively, while the majority class baselines for three tasks are 76.9%, 33.5%, 61.4%. The baseline that used *Interpersonal Pronoun*, and keywords "race", "racism" (directly indicating the necessities of self-narrations of personally experienced racism) achieved an accuracy of 77%. We see that adding linguistic features was useful for predicting personal experiences of racial discrimination. Accuracies obtained for all three tasks were notably higher than by-chance baselines and added between 4% and 12% in accuracy over baselines.

Given adequate validity for the classifiers, we then applied them to measure the personal experiences of racial discrimination contained in the 3146 thread messages in our full dataset. Among them, 1603 threads were identified as describing authors' personal experiences of racism. We found African-Americans reported being discriminated against in 1076 out of 1603 threads; 126 threads were about discrimination against East Asians (e.g. Chinese, Japanese); 62 against Hispanic/Latino; and 35 threads against South Asians (e.g. Indian, Pakistani). Although these percentages reflect only the incidence of reported racism experiences in this forum, it is worth noting the high number of posts indicating personally experienced racism by African-Americans despite they are underrepresented on Reddit[2].

Over those 1603 messages, we trained a Latent Dirichlet Allocation (LDA) model to discover the hidden topics related to self-narration of racial discrimination. The model was set to derive 20 topics based on model perplexity and human judgment. Per common practice, we manually assigned a label to each topic. Example topics include *Taunting* (e.g. 'wrongly', 'crush', 'blah'), *Appearance and Viewing* (e.g. 'apparent', 'rejected', 'behaved', 'viewing'), *Effects of Racism on People* (e.g. 'assaulted', 'effort', 'awakened', 'asleep', 'tears'), *Accent and Behavior* (e.g. 'sounding', 'acted', 'shaken'), *Violence* (e.g. 'corner', 'violent', 'charge', 'awful'), *shooting* (e.g. 'shooting', 'shoot', 'truth'), *Sexism* (e.g. 'gender', 'ladies', 'haircut'), etc. These reveal different dimensions of racism victims' experiences, from why and how they were discriminated against to what negative effects they had to bear as the result of racism. Note that these topics can overlap. For instance, comments on a person's accent could take the form of taunting, or suffering violence can generate an effect of fear.

Characterizing Support

Posts self-narrating an experience of racial discrimination may elicit supportive responses from the community. To understand the nature of community replies, we designed a predictive model to classify which type of support is provided to the victims of racial discrimination. To take into account the uniqueness of racial discrimination, we manually went through the responses from the community and came up with the following set of finer-grained emotionally supportive actions (Cohen 2004): (1) **Commiseration** refers to others sharing similar personal experience. (2) **Encouragement** gives courage to others or expresses hope that situation will improve. (3) **Empathy** contains feeling sorry and/or understanding and feeling what another person is experiencing. Finally, we consider (4) **Informational Support** to be support that provides information or advice.

Given a message about self-narration of personal experiences of racism, and a reply to it, we asked Turkers to judge which types of support are provided in the reply. We also provided **Other Support** to capture supportive actions that are not listed above, and **No support** to refer to no provision of supportive actions. We annotated 1603 message-reply pairs using the same criteria to select Turkers as in the racism post annotation task. Each pair was labeled by five different Turkers, who received $0.08 for rating each message-reply pair. We acquired an overall ICC score of 0.777. We used a majority voting scheme to assign labels to a post. After removing replies lacking agreement, we obtained 1465 messages that had sufficiently consistent ratings across different workers. We designed five classifiers to predict different supportive actions contained in replies. We used the same set of features as the identification of experiences of racism, and trained Logistic Regression models over the labeled dataset, as presented in Table 2. Given that these models achieved adequate levels of accuracy around 80%, we then applied them to all 5212 replies to better understand the nature of support on Reddit.

Which Support Comes First? Among the 1603 thread messages that describe personal experiences of racial discrimination, 682 threads received replies from the community. We analyzed the first reply of each thread and computed how many threads received empathy in their first replies, how many received commiseration first, etc. We found that a majority of threads (58%) have encouragement in the initial reply, while only 22% of threads contain informational sup-

[2] http://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

Variables	Model 1	Model 2
Time Elapsed	-0.042***	-0.070***
Word Count	0.180***	0.102***
Encouragement		-0.004
Informational Support		-0.041**
Empathy		0.007
Commiseration		0.141***

Table 3: Prediction of Usefulness of Responses. Here, ***: p<0.001; **:p<0.01; *:p<0.05; .:p<0.1. N = 5211.

port in the initial reply, 13% received commiseration first, and only 8% threads received empathy in their initial replies.

What Support is Most Appreciated by the Community? To investigate this, we examine how different support types relate to their collective appraisal as measured by their Reddit score. The Reddit score[3] for a post is the number of upvotes from other members it receives minus its number of downvotes, which reflects a common appraisal by a community of Reddit subscribers. We designed two linear regression models to predict the score[4] of each response based on the types of support it received. We considered the estimated probability of different types of support as input features together with control variables such as time elapsed and word count (the total number of words in a reply). Time Elapsed is defined as the number of days elapsed after the initial post was made. Model 1 describes the effect of the control variables, which indicate that more immediate and longer response posts are more valuable (Table 3). In Model 2, we see that commiseration was the most appreciated support type by the community, which might trigger thread starters' social comparison, and further increase their motivation and hope (Taylor and Lobel 1989). Notably, informational support yielded negative weights after including our controls, suggesting that information exchanged may lack accuracy or credibility compared to other professional sites (Wang, Kraut, and Levine 2012).

Conclusion and Discussion

We conducted a large scale analysis of social media posts and replies to identify self-narrations of racial discrimination and corresponding community support. Specifically, we developed a set of machine learning models to detect instances of personally experienced racial discrimination, and then identified which types of support (empathy, encouragement, commiseration and informational support) were provided in subsequent replies, with reasonable accuracy. Our hope then, is that even though racism is still prevalent, we can leverage these personal accounts as tools for both social scientific understanding at population scale of the nature of racism being experienced, and for training data to build systems for identification of self-narration of racism and corresponding social support. This work has several limitations. First, our measures of self-narrations of racial discrimination are judged by Turkers who themselves may not be victims of racism or even who might have experienced racism and thus down-rate messages. Second, self-reported satisfaction scores of thread starters towards received replies are not available in our corpus, and we urge future research to utilize direct feedback from thread starters to validate and extend our findings on which support are most appropriate to resolve their distress.

References

Bartlett, J.; Reffin, J.; Rumball, N.; and Williamson, S. 2014. Anti-social media. *Demos* 1–51.

Chaudhry, I. 2015. # hashtagging hate: Using twitter to track racism online. *First Monday* 20(2).

Cicchetti, D. V. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 6(4):284.

Clark, R.; Anderson, N. B.; Clark, V. R.; and Williams, D. R. 1999. Racism as a stressor for african americans: A biopsychosocial model. *American Psychologist* 54(10):805.

Cohen, S. 2004. Social relationships and health. *American Psychologist* 59(8):676.

De Choudhury, M.; Jhaver, S.; Sugar, B.; and Weber, I. 2016. Social media participation in an activist movement for racial equality. In *ICWSM*.

Mays, V. M.; Cochran, S. D.; and Barnes, N. W. 2007. Race, race-based discrimination, and health outcomes among african americans. *Annual Review of Psychology* 58:201.

Melican, D. B., and Dixon, T. L. 2008. News on the net credibility, selective exposure, and racial prejudice. *Communication Research* 35(2):151–168.

Olteanu, A.; Weber, I.; and Gatica-Perez, D. 2015. Characterizing the demographics behind the# blacklivesmatter movement. *arXiv preprint arXiv:1512.05671*.

Paradies, Y.; Priest, N.; Ben, J.; Truong, M.; Gupta, A.; Pieterse, A.; Kelaher, M.; and Gee, G. 2013. Racism as a determinant of health: a protocol for conducting a systematic review and meta-analysis. *Systematic Reviews* 2(1):1.

Schuman, H. 1997. *Racial attitudes in America: Trends and interpretations*. Harvard University Press.

Seawell, A. H.; Cutrona, C. E.; and Russell, D. W. 2012. The effects of general social support and social support for racial discrimination on african american womens well-being. *Journal of Black Psychology* 0095798412469227.

Taylor, S. E., and Lobel, M. 1989. Social comparison activity under threat: Downward evaluation and upward contacts. *Psychological review* 96(4):569.

Wang, Y.-C.; Kraut, R.; and Levine, J. M. 2012. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. In *CSCW*, 833–842. ACM.

[3] https://www.reddit.com/wiki/faq

[4] We acknowledge that this helpfulness of responses are rated from the perspective of the community members, not the victims themselves. We urge readers to interpret our findings with caution.